Lancaster University Management School

Economics Working Paper Series

2015/022

Environmental agreements under asymmetric Information

Aurélie Slechten

The Department of Economics
Lancaster University Management School
Lancaster LA1 4YX
UK

Environmental agreements under asymmetric information[*]

Aurélie Slechten[†]

Abstract

In a two-country model, I analyze international environmental agreements when a country's emission abatement costs are private information and participation to an agreement is voluntary. I show that the presence of asymmetric information may prevent countries from reaching a first-best agreement if this information asymmetry is too high. I propose a new channel to restore the feasibility of the first-best agreement: pre-play communication. By revealing its abatement cost through a certification agency in a pre-play communication stage, a country commits not to misreport this abatement cost during the negotiations of an agreement. Hence, following certification by at least one country, information asymmetry is reduced. Certification restores the feasibility of the first-best agreement except for intermediate levels of information asymmetry. For those levels, one country undergoing certification is not always sufficient to restore the feasibility of the first-best but it is impossible to find transfers between countries such that they both optimally accept to undergo certification. One country has always an incentive to free-ride on the other country's certification.

Keywords: environmental agreements; asymmetric information; certification; information disclosure

JEL Codes: Q54, D82

[*] This research has received funding from the FRS - FNRS. I am especially grateful to Estelle Cantillon, for her very helpful comments and constructive discussions. I am also indebted to Nicolas Sahuguet, Patrick Legros, Paola Conconi, Andreas Lange and David Martimort for their comments. I also thank the participants of the internal seminar at ECARES, the 11th Journées André-Louis Gérard Varet, the 21th Annual Conference of the EAERE, the 2013 ENTER Jamboree and the 2014 Conference on Auctions, Competition, Regulation and Public Policy.

[†] Lancaster University Management School, a.slechten@lancaster.ac.uk

1 Introduction

The classical explanation for the failure of international negotiations on environmental issues is the free-rider problem: countries have the possibility to opt out of the negotiations while still enjoying the benefits of the global agreement. It is well-known from the mechanism design literature that inefficiencies arising from the free-rider problem are particularly relevant in contexts plagued by information asymmetry. For example, Rob (1989), and Mailath and Postlewaite (1990) stress the role of participation constraints to generate inefficiency. However, international environmental agreements are generally preceded by discussion rounds during which concerned parties do not negotiate pollution abatement targets but can communicate with each other and exchange information. This paper takes a mechanism design approach and studies the effect of asymmetric information about pollution abatement costs on the feasibility of an efficient environmental agreement when participation to this agreement is voluntary and when countries can communicate before negotiations start.

In the case of international environmental agreements, information asymmetry can also be a problem for negotiating this type of agreement. Information asymmetry can be understood as a lack of real knowledge about abatement options and costs available in other countries. For example, Espinola-Arredondo and Munoz-Garcia (2012) suggest that, in the context of climate change, the penetration rate of clean technologies along all industries in a specific country, and so the abatement cost of this country, is difficult to observe by outsiders. Information asymmetry can also be interpreted in a broader sense as the political cost necessary to implement a certain level of emission abatement (i.e. countries may have private information about the relative weight of environmental problems in governments' agenda).[1] Whatever the source of information asymmetry, this will be an issue for the negotiations of abatement targets. Countries have an incentive to exaggerate their privately known pollution abatement cost (or understate their privately known abatement benefit) in order to reduce the effort they have to supply and leave most of the burden of abatement on other countries.

Another characteristic of international environmental agreement is the fact that international cooperation develops overtime (see Wagner, 2001). This development usually follows a particular pattern: countries first agree on an initial agree-

[1] See for example Konrad and Thum (2014).

ment, i.e. an *umbrella convention*, that generally does not contain any countries' emission reduction targets but sets up the institutions entitled to gather information transmitted by countries and to negotiate all subsequent emission reduction targets. In a second stage, countries negotiate an agreement with emission reduction targets and financial or technology transfers.

In the context of climate change negotiations, countries first agreed on the United Nation Framework Convention on Climate Change that in particular established the Conference of Parties as a supreme body entitled to negotiate all subsequent protocols and amendments. By signing the UNFCCC, industrialized countries committed themselves to provide the Conference of the Parties with clear data about their greenhouse gas (GHG) emissions and about regional programmes containing measures to mitigate climate change and with information related to implementation (which could give an indication of the political willingness to implement emission reductions). A subsidiary body of the UNFCCC was in charge of assessing this information, which was thus at least partially verifiable. In a second stage, countries negotiated the Kyoto Protocol. Other examples are the Convention of Long-Range Transboundary Air-Pollution (LRTAP) or the Vienna Convention for the Protection of the Ozone Layer (which preceded the Montreal protocol). Under the LRTAP Convention, the concerned parties set up an emission monitoring system under the auspices of the United Nations Economic Commission for Europe before negotiating the first protocol with emission reduction targets. The Vienna Convention established the UNEP as a secretariat and asked this body to convene workshops to develop a more common understanding of factors affecting the ozone layer including costs and effects of possible control measures (see Benedick, 1998).

As shown by these examples, in many cases, there exists an international agency that collects information transmitted by countries. If the umbrella Convention setting up this agency is designed in such a way that it is allowed to verify countries' information (for example, by sending experts), this agency can to some extent be used as a certification device. The contribution of this paper is twofold. First, I show how the presence of information asymmetry about abatement costs may exacerbate the free-rider problem.[2] The second contribution of this paper is the introduction of a communication stage during which countries have the possibility to exchange verifiable information trough an international agency. In doing

[2] Here I assume information asymmetry about abatement costs, but we can also build a model in which information asymmetry is about abatement benefits.

so, I also propose a new channel to restore the feasibility of an efficient agreement: pre-play communication with certification.

My analysis is carried out within the framework of the private provision of a public good under asymmetric information. The public good considered here is the reduction of some pollutant. To provide this public good, countries incur emission abatement costs. Each country knows its own abatement cost, but not that of the other country. An environmental agreement consists of binding commitments to some emission abatement levels and monetary transfers between countries. The efficient or first-best agreement is the one that maximizes global welfare. Participation to this first-best agreement is voluntary. This agreement is feasible if all countries are willing to participate to it and if they all reveal their abatement cost truthfully. The objective of the model is to analyze the conditions on the economic environment (preferences, distribution of abatement costs) for which the first-best agreement is feasible.

I first consider a model without pre-play communication. Due to a trade-off between ensuring participation to the agreement and ensuring truth-telling, a first-best agreement is not feasible when the range of the distribution of abatement costs is too large (in other words, when the information asymmetry is high).[3]

Then, I introduce a pre-play communication stage, i.e. an umbrella convention that establishes an international agency entitled to gather information about privately known abatement costs provided by participating countries. In this model, I assume that this information is totally verifiable: the international agency can monitor and certify the countries' privately known information if sovereign countries give their consent.

The effects of certification are twofold. On the one hand, the information asymmetry between countries is reduced. On the other hand, the country certifying its abatement cost loses the possibility to misreport this abatement cost, and thereby may see its monetary transfer in the first-best agreement reduced.

Compared to a model without the possibility of certification, I show that there exist three types of equilibrium in the two-stage game. First, for low levels of information asymmetry, there is always one country using the certification agency as

[3] Similar inefficiencies were pointed out in related setups. In the context of public good economies, Laffont and Maskin (1979) have shown that no truthful and efficient mechanisms may exist if individual rationality constraints are taken into account. In the private goods case, Myerson and Satterthwaite (1983) showed the impossibility of attaining ex-post efficiency with an incentive-compatible and individually rational mechanism.

a commitment device to truthfully disclose its abatement cost. The induced reduction of information asymmetry allows countries to reach the first-best agreement for all realizations of abatement costs. Second, for intermediate levels of information asymmetry, there is still one country using the certification agency, but the first-best is not implementable for all realizations of abatement costs. Finally, when the information asymmetry is very high, it is possible to find transfer schemes such that both countries decide to reveal their type through the certification agency before negotiations. In this case, the first-best is achieved for all realizations of abatement costs. Due to this high information asymmetry, the probability that only one certification is sufficient to reach the first-best is very low, so countries do not have an incentive to free-ride on each others' certification. Therefore, even if countries have the possibility to get rid of information asymmetry and implement the first-best agreement, there still exist economic environments for which a first-best agreement is not feasible. This is due to the fact that it is impossible to design transfers that avoid free-riding by one country in the use of certification.

Some papers have developed specific applications of the mechanism design theory to environmental economics (e.g. Rob, 1989; Baliga and Maskin, 2003; Caparros et al., 2004; Konrad and Thum, 2014 or Helm and Wirl, 2015). The model the closest to the one developed in this paper is that of Martimort and Sand-Zantman (2015). They also highlight a trade-off between solving free riding due to asymmetric information and due to voluntary participation and analyze the characteristics of a second-best mechanism. They show that the optimal mechanism admits a simple approximation by menus. By contrast, my paper proposes a channel to restore the feasibility of the first-best agreement, i.e. pre-play communication with certification.

There exists an extensive literature on verifiable communication in the context of mechanism design. Some papers focus on the question of identifying an appropriate form of the Revelation principle (e.g. Forges and Koessler, 2005; Bull and Watson, 2004; Bull and Watson, 2007; Deneckere and Severinov, 2008). More recently, Hagenbach et al. (2014) focus on the conditions to reach full disclosure of privately held information when the players can make pre-play certifiable statements. In another context, Benoît and Dubra (2006) slightly modify a variety of auction models by adding a preliminary stage in which one player can send a verifiable signal revealing his private information before the auction. They show that a player will reveal all of his information in equilibrium, even though this lowers his ex-ante payoff. In this paper, I focus on the private provision of a public good.[4]

The remainder of the paper is organized as follows. Section 2 lays out the main assumptions of the two-country model. Section 3 shows the effect of asymmetric information on the feasibility of the first-best agreement. In section 4, I introduce pre-play communication. I conclude in section 5.

2 The model

2.1 The setting

There are two countries, $i = 1, 2$, that exert some non-negative pollution abatement efforts a_i. Global abatement benefit is simply the total quantity of abatement, i.e. $(a_1 + a_2)$. Each country i receives a share of this global benefit: $b_i(a_1 + a_2)$, where $b_i > 0$ and $b_1 + b_2 = 1$.[5]

Countries are heterogeneous in terms of their marginal cost of abatement. By exerting abatement effort a_i, country i incurs a cost of $\frac{1}{2\theta_i}a_i^2$. For tractability, I adopt a quadratic form where θ_i can be interpreted as the characteristic of the technology of country i.[6] Finally, country i may receive a transfer t_i for undertaking the requested abatement. Country i's utility function is given by:

$$b_i(a_1 + a_2) - \frac{a_i^2}{2\theta_i} + t_i \tag{1}$$

The parameter θ_i is privately observed by country i. The types θ_i are independently drawn from the same uniform distribution defined on the support $[\underline{\theta}, \bar{\theta}]$, with $0 < \underline{\theta} < \bar{\theta}$. The cumulative and probability distribution function are respectively given by $F(\theta_i) = (\theta_i - \underline{\theta})/(\bar{\theta} - \underline{\theta})$ and $f(\theta_i) = 1/(\bar{\theta} - \underline{\theta})$. The mean is denoted by $E[\theta]$. The country with the highest θ_i is the most efficient at undertaking abatement efforts. I denote the set of states of the world as: $\Theta = [\underline{\theta}, \bar{\theta}] \times [\underline{\theta}, \bar{\theta}]$

For future reference, I define an *economic environment* as follows:

Definition 1 *An economic environment Ω consists of:*

[4] Note that in public good games, some papers have analyzed the role of pre-play *cheap talk*. See for example, Palfrey and Rosenthal (1991) and Agastya et al. (2007).

[5] This formulation is similar to McGinty (2007)

[6] This type of abatement cost has been used in other papers about international environmental agreements: Hoel and Schneider (1997), Kolstad (2005) or Martimort and Sand-Zantman (2015).

- *A range for the distribution of types* $\Delta\theta = (\bar{\theta} - \underline{\theta})$*;*

- *Individual marginal benefits from abatements* (b_1, b_2)*.*

In this setting, the first-best abatement levels are denoted by (a_1^{FB}, a_2^{FB}) and are defined as:

Definition 2 *The first-best abatement levels* $(a_1^{FB}, a_2^{FB}) = (\theta_1, \theta_2)$ *are the abatement levels that maximize the global welfare:*

$$(a_1^{FB}, a_2^{FB}) \in argmax_{a_1,a_2} \ (a_1 + a_2) - \frac{a_1^2}{2\theta_1} - \frac{a_2^2}{2\theta_2}$$

The most efficient countries in terms of abatement are those that abate the most at the first-best. In this paper, I will investigate whether these abatement levels can be implemented under asymmetric information about abatement costs and voluntary participation to an agreement.

2.2 Mechanisms

A mechanism is an agreement concluded between both countries and it consists of abatement levels a_i and transfers t_i for each country.[7] By the revelation principle, there is no loss of generality in restricting our attention to direct and truthful revelation mechanisms (Myerson, 1982). A direct revelation mechanism, $y = (a_1, a_2, t_1, t_2)$, is composed of a level of abatement for each country $a_i : \Theta \times \Theta \to \mathbb{R}$ that describes the abatement effort of each country as a function of countries' reported types, and a transfer $t_i : \Theta \times \Theta \to \mathbb{R}$ that describes each country's received transfer from undertaking the requested abatement effort as a function of countries' reported types.

I denote the utility of country i of type θ_i from the direct revelation mechanism $y(\hat{\theta}_1, \hat{\theta}_2)$, where $\hat{\theta}_i$ are the reported types for each country i, by:

$$V_i(y(\hat{\theta}_1, \hat{\theta}_2)|\theta_i) = b_i(a_1(\hat{\theta}_1, \hat{\theta}_2) + a_2(\hat{\theta}_1, \hat{\theta}_2)) - \frac{a_i^2(\hat{\theta}_1, \hat{\theta}_2)}{2\theta_i} + t_i(\hat{\theta}_1, \hat{\theta}_2)$$

Since in this paper, I am interested in the implementation of the first-best abatement levels (a_1^{FB}, a_2^{FB}), I define the first-best agreement as follows:

[7] Including transfers in environmental agreements is often explicit. Article 11 of the Kyoto Protocol allows for the possibility of monetary transfers from developed to developing countries through the Global Environment Facility.

Definition 3 *$y^{FB} = (a_1^{FB}, a_2^{FB}, t_1^{FB}, t_2^{FB})$ is the first-best agreement where $a_i^{FB} = \theta_i$ for $i = 1, 2$*

The expected utility of country i under this first-best agreement is given by:

$$V_i^{FB}(\theta_i) \equiv E_{\theta_j}[V_i(y^{FB}(\theta_1, \theta_2)|\theta_i)] = (b_i - \frac{1}{2})\theta_i + b_i E[\theta] + E_{\theta_j}[t_i^{FB}(\theta_1, \theta_2)] \qquad (2)$$

where $E_{\theta_j}[.]$ denotes the expectation over the possible types of country j.

To be implementable, a mechanism must satisfy three constraints. First, as countries are privately informed, the mechanism must be incentive compatible. Second, the mechanism must ensure that both countries want to join the agreement (i.e. individual rationality). Finally, there is also a budget balance constraint. These constraints are detailed below.

Bayesian incentive compatibility. Bayesian incentive compatibility of the mechanism $y(.,.)$, implies that the expected utility of country i must satisfy:

$$E_{\theta_j}[V_i(y(\theta_i, \theta_j)|\theta_i)] = \max_{\hat{\theta}_i \in [\underline{\theta}, \bar{\theta}]} E_{\theta_j}[V_i(y(\hat{\theta}_i, \theta_j)|\theta_i)] \qquad (3)$$

In other words, truth-telling gives country i the highest possible expected utility, provided the other country j does. By pretending to be a little bit less efficient in terms of abatement, i.e pretending to be of a type $\theta_i - \epsilon$, a country of type θ_i can abate at the same level as the less efficient type $\theta_i - \epsilon$ but at a lower marginal cost. The difference in terms of marginal cost is:

$$E_{\theta_j}\left[\frac{a_i^2(\theta_i - \epsilon, \theta_j)}{2(\theta_i - \epsilon)}\right] - E_{\theta_j}\left[\frac{a_i^2(\theta_i - \epsilon, \theta_j)}{2\theta_i}\right]$$

To ensure truth-telling, the mechanism must reward therefore the most efficient types by an extra amount,

$$E_{\theta_j}[V_i(y(\theta_i, \theta_j)|\theta_i) - V_i(y(\theta_i - \epsilon, \theta_j)|\theta_i)]$$

that is equal to this difference in terms of marginal cost. By letting ϵ tend to zero, we get the result of lemma 1:

Lemma 1 *The direct revelation mechanism $y(.,.)$ satisfies Bayesian incentive compatibility if and only if $E_{\theta_j}[a_i^2(\theta_i, \theta_j)]$ is weakly increasing in θ_i and*

$$E_{\theta_j}[\dot{V}_i(y(\theta_i, \theta_j)|\theta_i)] = \frac{E_{\theta_j}[a_i^2(\theta_i, \theta_j)]}{2\theta_i^2} \qquad (4)$$

Where $\dot{V}_i(.)$ is the derivative of the utility function with respect to the announcement of country i.

It immediately follows that an incentive compatible mechanism must give a greater payoff to the most efficient countries. As, in the sequel, I focus on incentive compatible direct revelation mechanism, I will simplify the notation of the expected utility:

$$E_{\theta_j}[V_i(y|\theta_i)] \equiv E_{\theta_j}[V_i(y(\theta_i,\theta_j)|\theta_i)]$$

This is the expected utility of country i from the direct and truthful revelation mechanism $y(\theta_i,\theta_j)$ when this country i is of type $\theta_i \in [\underline{\theta},\bar{\theta}]$.

Budget balance. I assume that no external source of funds is available and that there is no waste of resource.[8] Hence, the ex-ante budget-balance constraint implies that:

$$E_{\theta_1\theta_2}[t_1(\theta_1,\theta_2)+t_2(\theta_1,\theta_2)] = 0 \tag{5}$$

where $E_{\theta_1\theta_2}$ denotes the expectation over the types of countries 1 and 2. The interpretation of this constraint is that the overall surplus generated by countries' abatement efforts should be equal to their overall payoff. Note that there is no loss of generality in using the ex-ante budget balance constraint instead of the more natural ex-post budget balance constraint $(t_1(\theta_1,\theta_2)+t_2(\theta_1,\theta_2)=0)$ because, following Börgers and Norman (2009), if types are independent, for every ex-ante budget-balanced mechanism, there exists an ex-post budget-balanced mechanism such that the allocation rule is unchanged and the interim expected payments are unchanged for all agents.

Individual rationality. Finally, participation to an environmental agreement y is voluntary. The outside option is the non-cooperative equilibrium. It consists for country i in choosing its abatement level to maximize its own utility. Since countries choose their abatement non-cooperatively, there is no transfer, i.e. $t_1 = t_2 = 0$.

Definition 4 *$y^N = (a_1^N, a_2^N, t_1^N, t_2^N)$ is the non-cooperative equilibrium if and only if*

- $a_i^N = \theta_i b_i = argmax_{a_i}\ b_i(a_i + a_j^N) - \frac{a_i^2}{2\theta_i} + t_i^N$
- $t_1^N = t_2^N = 0$

[8] I can relax this budget balance constraint, i.e. by assuming linkages with agreements in other areas (e.g. low tariffs conditional on ratifying the environmental agreement) but the main results would remain unchanged.

Countries do not internalize the benefit of their abatement level on the other country and there is an under-provision of emission abatements. Note that a_i^N is a dominant strategy (i.e. the abatement chosen is the same whatever the behavior of the other country).The expected utility of country i under the outside option is denoted:

$$V_i^N(\theta_i) \equiv E_{\theta_j}[V_i(y^N|\theta_i)] = \frac{b_i^2}{2}\theta_i + b_1 b_2 E[\theta] \tag{6}$$

Since countries know their type when deciding to join a treaty, the interim individual rationality constraint requires that:

$$E_{\theta_j}[V_i(y|\theta_i)] \geq V_i^N(\theta_i) \quad \text{for } i = 1, 2 \tag{7}$$

3 A model without pre-play communication

3.1 Complete information benchmark

As a benchmark, I will first assume that countries' types θ_i for $i = 1, 2$, are public knowledge, so that abatement and transfer levels do not have to be incentive compatible. However, the first-best mechanism must still satisfy budget balance and individual rationality.

Proposition 1 *In all economic environments Ω, the first-best agreement y^{FB} is implementable for all $\theta_i \in [\underline{\theta}, \bar{\theta}]$ $(i = 1, 2)$.*

Proof Using equations (2) and (7), the transfers necessary to make each country willing to participate to the first-best agreement are given by:

$$t_1(\theta_1, \theta_2) \;\geq\; \theta_1 \frac{b_2^2}{2} - \theta_2 b_1^2 \tag{8}$$

$$t_2(\theta_1, \theta_2) \;\geq\; \theta_2 \frac{b_1^2}{2} - \theta_1 b_2^2 \tag{9}$$

Using the budget-balance constraint $t_1(\theta_1, \theta_2) + t_2(\theta_1, \theta_2) = 0$, the condition for the implementation of a first-best agreement under complete information is the following:

$$\theta_1 \frac{b_2^2}{2} - \theta_2 b_1^2 \leq t_1(\theta_1, \theta_2) \leq \theta_1 b_2^2 - \theta_2 \frac{b_1^2}{2} \tag{10}$$

In equation (10), $\theta_1 \frac{b_2^2}{2} - \theta_2 b_1^2 < \theta_1 b_2^2 - \theta_2 \frac{b_1^2}{2}$ for all $\theta_i > 0$. Indeed, rearranging this inequality, we get:

$$-\theta_1 \frac{b_2^2}{2} < \theta_2 \frac{b_1^2}{2}$$

The result follows since the term on the left hand side is always negative, while the term on the right hand side is always positive. □

3.2 Two-sided asymmetric information

Now, I assume that the countries' types θ_i are private information. Combining interim individual rationality (7), bayesian incentive compatibility (4) and budget balance (5) yields the following proposition:

Proposition 2 *Given an economic environment Ω, the first-best agreement y^{FB} is implementable $\forall \theta_i \in [\underline{\theta}, \bar{\theta}]$ $(i = 1, 2)$ if and only if this economic environment satisfies:*

$$\frac{\Delta\theta}{\underline{\theta}} \leq \frac{b_1^2 + b_2^2}{2b_1 b_2} = L^0 \tag{11}$$

Proof See appendix A. □

The left hand side of condition (11) is the relative range of the distribution of types $\Delta\theta/\underline{\theta}$. By lemma 1, we know that to ensure truth-telling the most efficient countries must be given the highest payoffs. If information asymmetry, measured by the range $\Delta\theta$, is high, the incentives of the most efficient types for misreporting are big. Avoiding such free-riding requires large compensations, i.e. large information rent, for those very efficient countries.

The right hand side of condition (11) depends on countries' asymmetry in terms of marginal benefit from abatement. Specifically, L^0 is the lowest when countries are symmetric, i.e. when $b_1 = b_2 = 0.5$. To interpret this result, I define the expected collective gains from reaching the first-best agreement (or the efficiency gains) as the difference in expected welfare between the first-best agreement W^{FB} and the non-cooperative equilibrium W^N:

$$E_{\theta_1,\theta_2}[W^{FB} - W^N] = E_{\theta_1,\theta_2}[V_1^{FB}(\theta_1) + V_2^{FB}(\theta_2)] - E_{\theta_1,\theta_2}[V_1^N(\theta_1) + V_2^N(\theta_2)]$$

With the ex-ante budget balance constraint (5), these efficiency gains must be fully redistributed among countries. Using equations (2) and (6) yields:

$$E_{\theta_1,\theta_2}[W^{FB} - W^N] = \left(\frac{b_1^2}{2} + \frac{b_2^2}{2}\right)(E[\theta] + E[\theta]) = (b_1^2 + b_2^2)\frac{\bar{\theta} + \underline{\theta}}{2}$$

When countries are symmetric, the non-cooperative equilibrium welfare is closer to the first-best welfare. the efficiency gains that can be redistributed among countries to ensure participation and truth-telling are lower. These expected gains are the lowest for symmetric countries.

In other words, condition (11) states that efficiency gains must be higher than information costs (measured by the level of information asymmetry). To understand this result, we need to figure out the impact of the economic environment Ω on individual rationality and incentive compatibility.

As the budget balance constraint must always be satisfied, the compensations granted to the most efficient types are limited by the necessity to ensure participation of all types, including the least efficient ones. When the range of the distribution of types is very large or when efficiency gains are low, one cannot find incentive compatible transfers that implement the first-best abatement levels and that give all types (including the least efficient ones) strictly more than their expected non-cooperative payoffs. For economic environments that does not satisfy condition (11), we can say that there is a tension between incentive compatibility, budget-balance and individual rationality that prevents countries from reaching the first-best agreement. Compared to the benchmark case, the presence of asymmetric information reduces the set of economic environments for which countries can implement the first-best agreement.

Transfers under the first-best agreement. When the economic environment satisfies condition (11), the first-best abatement levels $a_i^{FB}(\theta_1, \theta_2) = \theta_i$ can be implemented using transfers that satisfy Bayesian incentive compatibility, individual rationality and budget balance. We have shown that if a mechanism is bayesian incentive compatible, the expected utility of each country i must satisfy equation (4). Integrating this equation yields:

$$E_{\theta_j}[V_i(y|\theta_i)] = V_i(\underline{\theta}) + \int_{\underline{\theta}}^{\theta_i} \frac{E_{\theta_j}[a_i^2(s, \theta_j)]}{2s^2} ds$$

where $V_i(\underline{\theta}) = E_{\theta_j}[V_i(y|\underline{\theta})]$. Using equation (1) and rearranging terms, we can recover the transfers under the first-best agreement:

$$E_{\theta_j}[t_i^{FB}(\theta_i, \theta_j)] = t_i^{FB}(\theta_i) = (a_i^{FB}(\theta_i) - a_i^N(\theta_i)) - (a_i^{FB}(\underline{\theta}) - a_i^N(\underline{\theta})) + t_i^{FB}(\underline{\theta})$$

where $a_i(\theta_i) = E_{\theta_j}[a_i(\theta_i, \theta_j)]$ is the expected abatement level and $t_i(\theta_i)$ is the expected level of transfer.[9] This can also be rewritten as:

$$t_i^{FB}(\theta_i) = (1 - b_i)(\theta_i - \underline{\theta}) + t_i^{FB}(\underline{\theta})$$

This expected first-best transfer is increasing in country i's type. The higher the type θ_i, the higher the term $(a_i^{FB}(\theta_i) - a_i^N(\theta_i))$, i.e. the more the country should

[9] When deciding to join or not an agreement and to tell their true type or not, countries only know their private type θ_i.

abate in the first-best agreement. It is then necessary to give efficient countries a higher transfer to give them an incentive to reveal truthfully their type. This transfer negatively depends on b_i: the higher the marginal benefit from abatement, the higher the contribution of the country to the mechanism. This makes sense: if a country cares a lot about the environment, it will be ready to contribute more in order to implement the first-best agreement.

This transfer also depends on $t_i^{FB}(\underline{\theta})$, which is the first-best transfer for the least efficient type. This transfer must satisfy the individual rationality and I show in appendix A, that if it is the case, then individual rationality is also satisfied for more efficient types.

3.3 One-sided asymmetric information

Finally, as it will be useful in the next section, I also derive the condition under which the first-best agreement is implemented when only one country has private information. Considering that the resolution of this problem is very similar to the previous case, I just report the main results. Details can be found in Appendix B.

Assume that country i's type is public knowledge and country j's type is privately known. I first derive the following Lemma, which states that, given the economic environment, the first-best agreement can be implemented if country i is sufficiently efficient in terms of abatement.

Lemma 2 *When country i's type is public knowledge and country j's type is privately known, the first-best agreement y^{FB} is implementable $\forall \theta_j \in [\underline{\theta}, \bar{\theta}]$ if and only if $\theta_i \geq \tilde{\theta}_i$ with*

$$\tilde{\theta}_i = \frac{b_i}{b_j}(\bar{\theta} - \underline{\theta}) - \frac{b_i^2}{b_j^2}\underline{\theta} \tag{12}$$

Proof See Appendix B. □

As before, the intuition behind this result relies on the efficiency gains:

$$E_{\theta_j}[W^{FB} - W^N] = \frac{b_i^2}{2}E[\theta] + \frac{b_j^2}{2}\theta_i$$

These gains are increasing in the publicly known type θ_i. The more efficient in terms of abatement country i is, the larger the efficiency gains available to solve the tension between participation and truth-telling are.

The critical type $\tilde{\theta}_i$ is the threshold from which the first-best agreement can be implemented when country i's type is public knowledge. It is increasing in the relative range of the distribution $\frac{\Delta\theta}{\underline{\theta}}$: the tension between incentive compatibility, budget-balance and individual rationality is stronger for larger ranges of the distribution. A higher value of θ_i is then necessary to solve the tension.

If $\tilde{\theta}_i \leq \underline{\theta}$, the first-best agreement y^{FB} will be implementable for all possible publicly known types $\theta_i \in [\underline{\theta}, \bar{\theta}]$. This will be the case if the economic environment satisfies the following condition:

$$\frac{\Delta\theta}{\underline{\theta}} \leq \frac{b_1^2 + b_2^2}{b_1 b_2}$$

It is then possible to derive a proposition similar to Proposition 2 (see Appendix B):

Proposition 3 *Assume that country i's type is public knowledge and country j's type is privately known. Given an economic environment Ω, the first-best agreement* y^{FB} *is implementable* $\forall \theta_i \in [\underline{\theta}, \bar{\theta}]$ *and* $\forall \theta_j \in [\underline{\theta}, \bar{\theta}]$ *if and only if this economic environment satisfies:*

$$\frac{\Delta\theta}{\underline{\theta}} \leq L^1 = \frac{b_1^2 + b_2^2}{b_1 b_2}$$

with $L^0 < L^1$.

4 A model with pre-play communication

Most of the existing environmental treaties (e.g. the Montreal Protocol, the Helsinki Protocol...) were preceded by an umbrella convention that set up the institutions under the auspice of which subsequent protocols were negotiated. This pattern of development implies that countries communicate before agreeing on pollution reduction targets. This communication generally takes place through an international agency (e.g. UNECE, UNEP...) that gathers information transmitted by countries.

In some cases, if countries allow it, the agency could monitor and certify this private information. Certification reduces the information asymmetry and constitutes a channel to restore the feasibility of the first-best agreement. Nevertheless, for this channel to be helpful in reaching the first-best agreement, countries must have incentives to undergo certification. The effect of certification is twofold. On the one hand, by revealing a country's type, certification reduces the information asymmetry, which was responsible for the tension between incentive compatibility, budget-balance and individual rationality constraints. On the other hand, certification implies the loss of the information rent for the country revealing its type.

Countries might thus have an incentive to free-ride on each others' certification (i.e. to stay privately informed) in order to keep their information rent.

To include the possibility of pre-play communication and certification, I modify the model presented in section 3 by considering a preliminary stage (stage 1 or pre-play communication stage) in which countries decide simultaneously whether or not to allow the international agency to monitor and certify their type. This action of certification is taken at an ex-ante stage, i.e. when countries do not know their own types.[10]

In stage 2, countries at least privately know their types and they negotiate an environmental agreement using the direct revelation mechanism that implements the first-best agreement if it is feasible and they resort to the non-cooperative outcome otherwise. This is the abatement stage. Note that the non-cooperative outcome y^N is independent of information, i.e. whether or not one country has revealed its type in stage 1. Indeed, transfers are equal to zero for all types and the non-cooperative abatement level only depends on each country's type ($a_i^N(\theta_i) = b_i\theta_i$). I solve this game backward.

Remark 1 *For economic environments satisfying $\frac{(\bar{\theta}-\underline{\theta})}{\underline{\theta}} \leq L^0$, the fact that countries undergo certifications does not help them to reach the first-best agreement: the first-best is implementable in stage 2 with or without certification (see proposition 2). The only effect of the possibility to certify its type is to change the first-best transfers. Independently of the certifications undertaken in stage 1, the equilibrium outcome of this two-stage game is that the first-best is always implemented. For this reason, I concentrate on economic environments satisfying $\frac{\Delta\theta}{\underline{\theta}} > L^0$ in which the first-best is not implementable without certification.*

4.1 Stage 2: abatement game

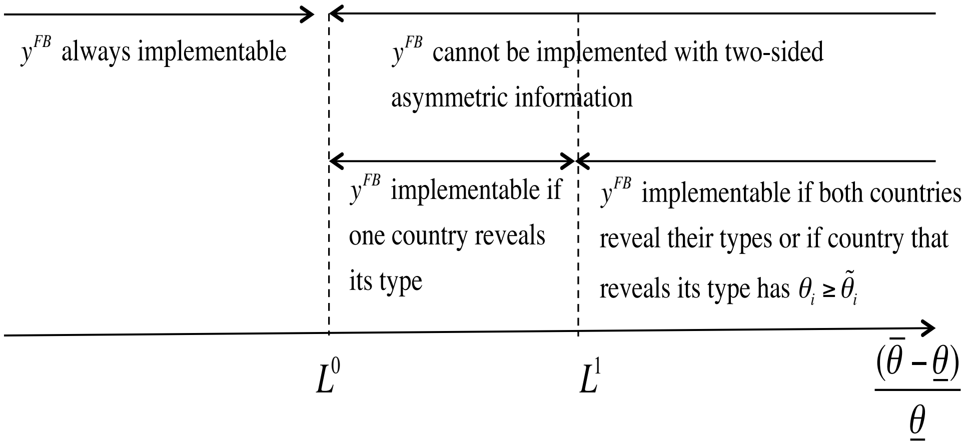
The equilibrium outcome in stage 2 depends on the actions taken in stage 1. If both countries have accepted transmitting their private information to the certification agency in stage 1, the equilibrium outcome in stage 2 is the first-best agreement y^{FB} (because there is complete information). If no country has accepted, then the equilibrium outcome in stage 2 is the non-cooperative equilibrium y^N.

[10] It would be more realistic to assume that countries decide whether or not to undertake certification at an interim stage (i.e. when each country knows its own abatement cost), but the resolution of the two-stage game would become intractable. This resolution would require looking at beliefs conditional on observable moves and non-equilibrium beliefs.

If only country i has allowed the international agency to certify its type θ_i, the equilibrium outcome in stage 2 depends on the economic environment and on the value of country i's type (see Lemma 2 and Proposition 3).

Figure 1 summarizes these findings for stage 2. For economic environment below the threshold L^1 the equilibrium outcome in stage 2 is for sure the first-best agreement y^{FB} if at least one country has allowed for certification in stage 1. For economic environments such that $\frac{\Delta\theta}{\underline{\theta}} \geq L^1$, the equilibrium outcome in stage 2 is the first-best agreement y^{FB} either if both countries have allowed for certification in stage 1 or if the only country that has given its consent in stage 1 turns out to have a type $\theta_i \geq \tilde{\theta}_i$. In other cases the equilibrium outcome in stage 2 is the non-cooperative equilibrium y^N.

Figure 1: Implementation of the first-best agreement depending on the economic environment



y^{FB} always implementable
y^{FB} cannot be implemented with two-sided asymmetric information
y^{FB} implementable if one country reveals its type
y^{FB} implementable if both countries reveal their types or if country that reveals its type has $\theta_i \geq \tilde{\theta}_i$
L^0
L^1
$\frac{(\bar{\theta}-\underline{\theta})}{\underline{\theta}}$

4.2 Stage 1: pre-play communication

Countries simultaneously choose whether they allow the international agency to certify their type (i.e. they agree on the role and the prerogatives of the international agency). Let s_i denote the action (or strategy) of country i in stage 1. The set of possible strategies is $S = \{C,\ NC\}$ for each i, where C means that country i allows for certification and NC means that country i refuses it. To find the equilibrium strategies of this game, I analyse countries' best responses.[11]

[11] I only look at pure strategies.

Best responses of country $\boldsymbol{i}$

If $s_j = NC$*, i.e. if country* j *refuses certification by the international agency*, the best response of country i is to allow for it for all economic environments satisfying $\frac{\Delta\theta}{\underline{\theta}} > L^0$. Indeed, by certifying its type, country i increases the probability that the first-best agreement is implemented in stage 2: this probability increases from zero to $(1 - F(\tilde{\theta}_i))$, where $F(.)$ is the cumulative density function of the type θ_i. Since the first-best (which satisfies the individual rationality constraint) gives a higher expected utility than the non-cooperative equilibrium, country i will always prefer to reveal its type through the certification agency. An immediate consequence of this best response is that there will always be at least one country that allows certification of its type at the equilibrium of stage 1.

If $s_j = C$*, if country* j *allows the international agency to certify its type*, the best response of country i is more difficult to determine. Intuitively, the incentive for certification will depend on the economic environment Ω (which determines the gains from reaching an agreement and the probability that country j is sufficiently efficient in terms of abatement) and on the comparison between the transfers received in the first-best in the two sub-games (i.e. when both countries allow for certification and when only country j allows for certification).

Consider $y(\theta_1, \theta_2) = (a_1^{FB}, a_2^{FB}, t_1, t_2)$, the mechanism chosen to implement the first-best abatement levels under complete information and $y'(\theta_1, \theta_2) = (a_1^{FB}, a_2^{FB}, t'_1, t'_2)$, the mechanism chosen to implement the first-best if only country j has allowed for certification. The expected utilities of country i under each strategy (C and NC) are the following:

- If country i allows for certification in stage 1 ($s_i = C$), there is complete information in stage 2 and the first-best is always implementable using a mechanism $y(\theta_1, \theta_2)$. The expected utility of country i in stage 1 is then given by (see equation (2)):

  $$E_{\theta_1,\theta_2}\left[\left(\frac{b_i^2}{2} - \frac{b_j^2}{2}\right)\theta_i + (b_i^2 + b_1 b_2)\theta_j + t_i(\theta_1, \theta_2)\right]$$

  which is equivalent to:

  $$E_{\theta_1,\theta_2}\left[\left(\frac{b_i^2}{2}\theta_i + b_1 b_2 \theta_j\right) + \left(b_i^2\theta_j - \frac{b_j^2}{2}\theta_i + t_i(\theta_1, \theta_2)\right)\right] \tag{13}$$

  The first term of the right hand side of (13) is the expected utility under the non-cooperative equilibrium (see equation (6)). The second term is the extra gain from reaching the first-best agreement. Because individual rationality constraint is satisfied, this extra gain is always positive.

- If country i does not allow for certification in stage 1 ($s_i = NC$), the first-best is implementable using a mechanism $y'(\theta_1, \theta_2)$ if and only if $\theta_j \geq \tilde{\theta}_j$. Denote $(1 - F(\tilde{\theta}_j))$ the probability that $\theta_j \geq \tilde{\theta}_j$, where $F(.)$ is the cumulative density function of the type θ_j. The expected utility of country i is given by:

$$\begin{aligned}
&F(\tilde{\theta}_j)E_{\theta_1,\theta_2}\left[\frac{b_i^2}{2}\theta_i + b_1 b_2 \theta_j | \theta_j < \tilde{\theta}_j\right] \\
&+ (1 - F(\tilde{\theta}_j))E_{\theta_1,\theta_2}\left[\frac{b_i^2}{2}\theta_i - \frac{b_j^2}{2}\theta_i + (b_i^2 + b_1 b_2)\theta_j | \theta_j \geq \tilde{\theta}_j]\right] \\
&+ (1 - F(\tilde{\theta}_j))E_{\theta_1,\theta_2}[t'_i(\theta_1, \theta_2) | \theta_j \geq \tilde{\theta}_j]
\end{aligned}$$

Rearranging terms yields:

$$\begin{aligned}
&E_{\theta_1,\theta_2}\left[\frac{b_i^2}{2}\theta_i + b_1 b_2 \theta_j\right] \\
&+ (1 - F(\tilde{\theta}_j))E_{\theta_1,\theta_2}\left[\left(b_i^2\theta_j - \frac{b_j^2}{2}\theta_i + t'_i(\theta_1, \theta_2)\right) | \theta_j \geq \tilde{\theta}_j\right]
\end{aligned} \tag{14}$$

The first term on the right hand side of (14) is the same as in (13) (the expected utility under the non-cooperative equilibrium) and the second term is the expected extra gain from reaching the first-best agreement, which is implementable only if country j is sufficiently efficient.

Comparing expressions (13) and (14) confirms our intuition that the best response of country i when $s_j = C$ depends on (1) the transfers used to implement the first-best abatement levels in stage 2 (either with certification of both countries $E_{\theta_1,\theta_2}[t_i(\theta_1, \theta_2)]$ or with country j' s certification only $E_{\theta_1,\theta_2}[t'_i(\theta_1, \theta_2)|\theta_j \geq \tilde{\theta}_j]$), (2) the relative intensity of preferences over abatement (i.e. values of b_i and b_j), and (3) the probability that country j is sufficiently efficient $(1 - F(\tilde{\theta}_j))$. The two last factors are determined by the economic environment Ω.

There are potentially many transfers $E_{\theta_1,\theta_2}[t_i(\theta_1, \theta_2)]$ or $E_{\theta_1,\theta_2}[t'_i(\theta_1, \theta_2)|\theta_j \geq \tilde{\theta}_j]$) that implement the first-best agreement in each case. In order to find the best response of country i when country j allows for certification, I have the following assumption

Assumption 1 *Country i will accept the certification of its type if and only if this country can be sure to obtain a higher expected utility by acting in this way. Otherwise it prefers to stay privately informed.*

Given this assumption, the best response of country i to $s_j = C$ will be to allow for certification ($s_i = C$) if, given the economic environment Ω, there exist mechanisms with transfers $t_i(\theta_1, \theta_2)$, such that (13) > (14) for all $t'_i(\theta_1, \theta_2)$. If such transfers $t_i(\theta_1, \theta_2)$ do not exist, country i prefers to stay privately informed ($s_i = NC$).

The equilibrium in stage 1

It is now possible to characterize the equilibrium of the game in stage 1. Given the best responses detailed above, we have the following lemma:

Lemma 3 *For economic environments satisfying*

$$\frac{\Delta\theta}{\underline{\theta}} \geq L^0$$

there is at least one country allowing the international agency to certify its type at the equilibrium of stage 1.

Moreover, both countries i will give their consent to certification at the equilibrium in stage 1 if there exist second-stage mechanisms with transfers $t_i(\theta_1, \theta_2)$ (that implement the first-best under complete information) such that (13) > (14) for all $t'_i(\theta_1, \theta_2)$, $i = 1, 2$. In this case, for both i, the best response of country i is $s_i = C$ when country j's strategy is $s_j = C$. If such transfers $t_i(\theta_1, \theta_2)$ does not exist for both i, then there is one country that prefers to stay privately informed when the other country allows for certification.

Intuitively, it is easier to find such transfers $t_i(\theta_1, \theta_2)$ when $F(\tilde{\theta}_j)$ is high (or equivalently when $\tilde{\theta}_j$ is high). For high values of $F(\tilde{\theta}_j)$, the probability that the certification of country j's type allows to reach the first-best is low, so that country i has a greater incentive to also give its consent in stage 1 in order to guarantee that the first-best is implemented in stage 2 (with complete information). For low values of $F(\tilde{\theta}_j)$, country i may have an incentive to free-ride on the other country's certification in order to stay privately informed and enjoy an information rent if the first-best is achieved in stage 2. This is shown formally in Lemma 4:

Lemma 4 *The equilibrium in stage 1 is unique and is such that both countries allow for certification if and only if the economic environment satisfies:*

$$\frac{\tilde{\theta}_1 - \underline{\theta}}{\bar{\theta} - \underline{\theta}} + \frac{\tilde{\theta}_2 - \underline{\theta}}{\bar{\theta} - \underline{\theta}} \geq 1 \tag{15}$$

or equivalently, if and only if:

$$\frac{\Delta\theta}{\underline{\theta}} \geq \frac{(b_1^2 + b_2^2)^2}{b_1 b_2 (b_1^2 + b_2^2 - b_1 b_2)} = L^2 > L^1 \tag{16}$$

For economic environments that do not satisfy (16)*, the equilibrium in stage 1 is unique and is such that one country allows for certification while the other country stays privately informed. The identity of the country allowing for certification will depend on the transfers used to implement the first-best agreement in stage 2, i.e.* $t'_i(\theta_1, \theta_2)$ *and* $t_i(\theta_1, \theta_2)$.

Proof See Appendix C. □

The threshold L^2 achieves its minimum at the point $b_1 = b_2 = 0.5$ (i.e. when countries are symmetric). The intuition behind this result is the same as for the threshold L^0: the more symmetric countries are, the lower the efficiency gains are. It is easier to give incentives to countries to undergo certification (i.e. to design transfers such that they both allow for certification) if these countries are more asymmetric (in terms of marginal benefits) or equivalently if expected collective gains from cooperation are higher.

Remark 2 *Assume, for instance, that* $b_i < b_j$. *Then,* $\tilde{\theta}_i < \bar{\theta}$ *for all ranges of the distribution of types, while* $\tilde{\theta}_j$ *may be such that* $\tilde{\theta}_j \geq \bar{\theta}$*(i.e.* $F(\tilde{\theta}_j) = 1$*). This is the case for economic environments satisfying:*

$$\frac{\Delta\theta}{\underline{\theta}} \geq \frac{b_1^2 + b_2^2}{b_1(b_2 - b_1)} = L^3 > L^2$$

where L^3 *is obtained by substituting* $\bar{\theta}$ *to* $\tilde{\theta}_2$ *in equation* (12)*. Therefore, for ranges above* L^3*,* $F(\tilde{\theta}_j) = 1$ *and condition* (16) *is automatically satisfied.*

When $(1 - F(\tilde{\theta}_j)) = 0$, the best response of country i is always to accept certification in stage 1 for all $t_i(\theta_1, \theta_2)$ and all $t'_i(\theta_1, \theta_2)$ because country j is never sufficiently efficient for its certification to allow countries to reach the first-best agreement. Country i will thus never have an incentive to free-ride on certification of country j's type.

4.3 Solution of the two-stage game

Three cases must be distinguished:

Case 1: Economic environments satisfying $L^0 < \frac{\Delta\theta}{\underline{\theta}} \leq L^1$

In these economic environments, $\tilde{\theta}_i = \underline{\theta}$, so that the equilibrium outcome of this game is the first-best agreement for all realizations of types, as soon as one country allows for certification, which will always happen along the equilibrium path by lemma 3.

Case 2: Economic environments satisfying $L^1 < \Delta\theta/\underline{\theta} < L^2$

There is only one country allowing for certification in stage 1. Since $\tilde{\theta}_i > \underline{\theta}$, certification by only one country is not always sufficient to reach the first-best agreement in the second stage. The country that certifies its type must be sufficiently efficient in terms of abatement. As a result, depending on the type of the country undergoing certification, the equilibrium in stage 2 may be either the first-best agreement y^{FB} or the non-cooperative outcome y^N.

Case 3: Economic environments satisfying $\frac{\Delta\theta}{\underline{\theta}} \geq L^2$

For these economic environments, both countries allow for certification at the equilibrium of stage 1. The equilibrium outcome in stage 2 is thus the first-best agreement y^{FB}, for all realizations of types.

Consolidating cases 1-3

We can now state a proposition about the possibility to implement the first-best abatement levels for all realizations of types in this two-stage game (similarly to Proposition 2):

Proposition 4 *Given an economic environment Ω, the first-best agreement y^{FB} is implementable for all $\theta_i \in [\underline{\theta}, \bar{\theta}]$ in the game with pre-play certification if this economic environment satisfies:*

$$\frac{\Delta\theta}{\underline{\theta}} \leq \frac{b_1^2 + b_2^2}{b_1 b_2} = L^1 \quad \mathit{Or} \quad \frac{\Delta\theta}{\underline{\theta}} \geq \frac{(b_1^2 + b_2^2)^2}{b_1 b_2 (b_1^2 + b_2^2 - b_1 b_2)} = L^2$$

with $L^1 < L^2$

For economic environments satisfying $\frac{\Delta\theta}{\underline{\theta}} \in (L^1, L^2)$, there exist some realizations of types for which the first-best is not implementable.

Proposition 4 states that certification can restore the feasibility of the first-best agreement for all types in two cases: for high or sufficiently low levels of information asymmetry.

If information asymmetry is high (threshold L^2), there exist mechanisms such that both countries optimally undergo certification. No country has an incentive to free-ride on the other's action because the risk to implement the non-cooperative equilibrium (outside option) in stage 2 is substantial and so a country's expected benefits of keeping its information rent are very low.

For lower levels of information asymmetry, there is always one country that prefers to free-ride on the certification of the other country. Due to this free-rider problem, the first-best agreement will not be part of the equilibrium path for all types. Nevertheless, if information asymmetry is particularly low (threshold L^1), both countries are sufficiently efficient in terms of abatement to solve the tension between incentive compatibility, budget-balance and individual rationality. Free-riding in the certification stage is no longer a problem and the first-best is implemented in stage 2 for all types' realizations.

The set of ranges (L^1, L^2) where the first-best is not necessarily part of the equilibrium path for all types is the largest when $b_1 = b_2$ (countries are symmetric). This corresponds to the case in which the expected collective gains from the first-best agreement are the lowest. Clearly, certifications can more easily solve the tension between incentive compatibility, budget-balance and individual rationality if those efficiency gains are high.

5 Conclusion

This paper takes a mechanism design approach to study the effect of asymmetric information about abatement costs on the feasibility of an efficient environmental agreement when participation is voluntary. Due to the tension between incentive compatibility and participation, a first-best agreement cannot always be reached. Then, I introduce a new channel to restore efficiency: pre-play communication or the possibility for countries to disclose truthfully their type through a certification agency.

I show that it is possible to find transfer schemes between countries such that a certification agency is created and at least one country allows this agency to certify its type. Adding the possibility of certification restores the feasibility of the first-best agreement if the level of asymmetry is either high (certification of both countries' types) or low (certification of one country's type).

One illustrative example of the model developed in this paper could be the climate change negotiations, even if this problem is substantially more complex (e.g. due to aggregate uncertainty about the physical process causing climate change and its impact on human beings?). Climate negotiations started with the 1992 United Nations Framework Convention on Climate Change (UNFCCC). This is an umbrella convention without any emission reduction targets, or interpreted in the light of the model, a pre-play communication stage. The UNFCCC tries, in some sense, to reduce information asymmetry between parties by providing some information about countries' private costs and benefits of climate change. Indeed, by signing the UNFCCC, countries committed themselves to provide the Conference of the Parties with information, e.g. about their GHG emissions and their regional programs to mitigate climate change (which can be seen as an indication of the political willingness to implement climate change policies). However, this obligation covered only industrialized countries.

This paper suggests that before negotiating a new climate change agreement, it could be useful to reinforce the UNFCCC by requiring, for example, that all countries report their GHG emissions and their technologies to reduce these GHG to the secretariat of the UNFCCC that should have the possibility to verify, at least partially, this information. The model also shows that transfers between industrialized and developing countries will play a crucial role. These transfers can take the form of monetary side payments through the Global Environment Facility (see article 11 of the Kyoto Protocol). By choosing appropriate transfers derived from the mechanism design theory, countries may have an incentive to disclose their private information through an international certification agency established by the UNFCCC. In the light of the model developed in this paper, this can help countries to reach an efficient climate agreement.

The model uses a lot of simplifying assumptions to highlight the effects of pre-play communication. A first important extension could be to consider that abatement efforts are not totally observable, so that there is a problem of moral hazard during the implementation of the first-best agreement. Second, in this paper, I assume that information transmitted during the pre-play stage is verifiable and can be certified by an international agency. A more realistic assumption would be that this information is only partially verifiable, such that abatement costs are still privately known but the beliefs about these abatement costs are modified by the first-stage action of information transmission.

Appendices

Appendix A: two-sided asymmetric information

Proof of Lemma 1

First, assume that the mechanism implementing the agreement $y = (a_1, a_2, t_1, t_2)$ is Bayesian incentive compatible. From equation (3), we get that $E_{\theta_j}[V_i(y(\theta_i, \theta_j)|\theta_i)]$ is the maximum of concave functions of θ_i. It is thus concave, absolutely continuous and almost everywhere twice differentiable. As a consequence, if a mechanism is Bayesian incentive compatible, it follows that:

$$E_{\theta_j}[\dot{V}_i(y(\theta_i, \theta_j)|\theta_i)] = \frac{E_{\theta_j}[a_i^2(\theta_i, \theta_j)]}{2\theta_i^2} > 0$$

Moreover, take $\theta_i \geq \hat{\theta}_i$ and rewrite the expected indirect utilities for these two types:

$$\begin{aligned} E_{\theta_j}[V_i(y(\theta_i, \theta_j)|\theta_i)] &= E_{\theta_j}\left[t_i(\theta_i, \theta_j) + b_i(a_i(\theta_i, \theta_j) + a_j(\theta_i, \theta_j)) - \frac{a_i^2(\theta_i, \theta_j)}{2\theta_i}\right] \\ E_{\theta_j}[V_i(y(\hat{\theta}_i, \theta_j)|\hat{\theta}_i)] &= E_{\theta_j}\left[t_i(\hat{\theta}_i, \theta_j) + b_i(a_i(\hat{\theta}_i, \theta_j) + a_j(\hat{\theta}_i, \theta_j)) - \frac{a_i^2(\hat{\theta}_i, \theta_j)}{2\hat{\theta}_i}\right] \end{aligned}$$

From equation (3), we can also write the two following inequalities:

$$\begin{aligned} E_{\theta_j}[V_i(y(\theta_i, \theta_j)|\theta_i)] &\geq E_{\theta_j}\left[t_i(\hat{\theta}_i, \theta_j) + b_i(a_i(\hat{\theta}_i, \theta_j) + a_j(\hat{\theta}_i, \theta_j)) - \frac{a_i^2(\hat{\theta}_i, \theta_j)}{2\theta_i}\right] \\ E_{\theta_j}[V_i(y(\hat{\theta}_i, \theta_j)|\hat{\theta}_i)] &\geq E_{\theta_j}\left[t_i(\theta_i, \theta_j) + b_i(a_i(\theta_i, \theta_j) + a_j(\theta_i, \theta_j)) - \frac{a_i^2(\theta_i, \theta_j)}{2\hat{\theta}_i}\right] \end{aligned}$$

These two inequalities imply that

$$E_{\theta_j}[a_i^2(\theta_i, \theta_j)]\left[\frac{1}{2\hat{\theta}_i} - \frac{1}{2\theta_i}\right] \geq E_{\theta_j}[V_i(y(\theta_i, \theta_j)|\theta_i) - V_i(y(\hat{\theta}_i, \theta_j)|\hat{\theta}_i)] \geq E_{\theta_j}[a_i^2(\hat{\theta}_i, \theta_j)]\left[\frac{1}{2\hat{\theta}_i} - \frac{1}{2\theta_i}\right] \tag{17}$$

which implies $E_{\theta_j}[a_i^2(\theta_i, \theta_j)] \geq E_{\theta_j}[a_i^2(\hat{\theta}_i, \theta_j)]$ for $\theta_i \geq \hat{\theta}_i$.

Reciprocally, assume that (4) holds and that $E_{\theta_j}[a_i^2(\theta_i, \theta_j)]$ is weakly increasing in θ_i. Then, $E_{\theta_j}[V_i(y(\theta_i, \theta_j)|\theta_i)]$ can be written as:

$$E_{\theta_j}[V_i(y(\theta_i, \theta_j)|\theta_i)] = E_{\theta_j}[V_i(y(\underline{\theta}, \theta_j)|\underline{\theta})] + \int_{\underline{\theta}}^{\theta_i} \frac{E_{\theta_j}[a_i^2(s, \theta_j)]}{2s^2} ds$$

Thereby for $\theta_i \geq \hat{\theta}_i$,

$$E_{\theta_j}[V_i(y(\theta_i,\theta_j)|\theta_i) - V_i(y(\hat{\theta}_i,\theta_j)|\hat{\theta}_i)] = \int_{\hat{\theta}_i}^{\theta_i} \frac{E_{\theta_j}[a_i^2(s,\theta_{-i})]}{2s^2}ds$$

To have incentive compatibility, we need that condition (3) is satisfied, which is equivalent (see first part of the proof) to the inequality (17):

$$E_{\theta_j}[V_i(y(\theta_i,\theta_j)|\theta_i)] \geq E_{\theta_j}\left[V_i(y(\hat{\theta}_i,\theta_j)|\hat{\theta}_i) + a_i^2(\hat{\theta}_i,\theta_j)\left[\frac{1}{2\hat{\theta}_i} - \frac{1}{2\theta_i}\right]\right]$$

We need to check if

$$E_{\theta_j}[V_i(y(\theta_i,\theta_j)|\theta_i) - V_i(y(\hat{\theta}_i,\theta_j)|\hat{\theta}_i)] = \int_{\hat{\theta}_i}^{\theta_i} \frac{E_{\theta_j}[a_i^2(s,\theta_j)]}{2s^2}ds \geq E_{\theta_j}[a_i^2(\hat{\theta}_i,\theta_j)]\left[\frac{1}{2\hat{\theta}_i} - \frac{1}{2\theta_i}\right]$$

Or equivalently if

$$\int_{\hat{\theta}_i}^{\theta_i} \frac{E_{\theta_j}[a_i^2(s,\theta_j) - a_i^2(\hat{\theta}_i,\theta_j)]}{2s^2}ds \geq 0$$

This is the case since $E_{\theta_j}[a_i^2(\theta_i,\theta_j)]$ is weakly increasing in θ_i. □

Proof of Proposition 2

We need to delineate the conditions to have a budget-balanced, Bayesian incentive compatible and interim individually rational first-best agreement.

Step 1: The first step of the analysis consists in consolidating the Bayesian Incentive compatibility and the budget-balance constraints. By lemma 1, we know that $E_{\theta_j}[\dot{V}_i(y(\theta_i,\theta_j)|\theta_i)] \geq 0$. Therefore, the expected utility function is increasing in the country's type θ_i. Integrating (4) yields:

$$E_{\theta_j}[V_i(y(\theta_i,\theta_j|\theta_i)] = V_i(\underline{\theta}) + \int_{\underline{\theta}}^{\theta_i} \frac{E_{\theta_j}[a_i^2(s,\theta_j)]}{2s^2}ds \tag{18}$$

with $V_i(\underline{\theta}) = E_{\theta_j}[V_i(y(\underline{\theta},\theta_j)|\underline{\theta})]$. This is the expected utility of country i when Bayesian incentive compatibility constraint is satisfied. The expected total welfare when Bayesian incentive compatibility constraints are satisfied for both i is thus given by:

$$\begin{aligned} &E_{\theta_1\theta_2}[V_1(y(\theta_1,\theta_2)|\theta_1)] + V_2(y(\theta_1,\theta_2)|\theta_2)] \\ &\qquad = V_1(\underline{\theta}) + V_2(\underline{\theta}) + E_{\theta_1\theta_2}\left[\int_{\underline{\theta}}^{\theta_1} \frac{E_{\theta_2}[a_1^2(s,\theta_2)]}{2s^2}ds + \int_{\underline{\theta}}^{\theta_2} \frac{E_{\theta_1}[a_2^2(\theta_1,s)]}{2s^2}ds\right] \end{aligned} \tag{19}$$

Integrating (19) by parts, we finally obtain the following expression for the expected total welfare:

$$E_{\theta_1\theta_2}[V_1(y(\theta_1,\theta_2)|\theta_1)] + V_2(y(\theta_1,\theta_2)|\theta_2)] = V_1(\underline{\theta}) + V_2(\underline{\theta})$$
$$+ \int_{\underline{\theta}}^{\bar{\theta}} \left(\int_{\underline{\theta}}^{\theta_1} \frac{E_{\theta_2}[a_1^2(s,\theta_2)]}{2s^2} ds \right) f(\theta_1) d\theta_1 + \int_{\underline{\theta}}^{\bar{\theta}} \left(\int_{\underline{\theta}}^{\theta_2} \frac{E_{\theta_1}[a_2^2(\theta_1,s)]}{2s^2} ds \right) f(\theta_2) d\theta_2 \tag{20}$$

where $f(\theta_i) = 1/(\bar{\theta} - \underline{\theta})$ because of the uniform distribution of types.

The budget-balance constraint requires $E_{\theta_1\theta_2}[t_1(\theta_1,\theta_2) + t_2(\theta_1,\theta_2)] = 0$.[12] Remember that the utility function of country i of type θ_i when reported types are (θ_1, θ_2) is given by:

$$V_i(y(\theta_1,\theta_2)|\theta_i) = b_i(a_1(\theta_1,\theta_2) + a_2(\theta_1,\theta_2)) - \frac{a_i^2(\theta_1,\theta_2)}{2\theta_i} + t_i(\theta_1,\theta_2) \tag{21}$$

Consolidating equations (20) and (21) using the budget-balance constraint yields the following equation:

$$E_{\theta_1\theta_2}\left[b_1(a_1(\theta_1,\theta_2) + a_2(\theta_1,\theta_2)) - \frac{a_1^2(\theta_1,\theta_2)}{2\theta_1} + b_2(a_1(\theta_1,\theta_2) + a_2(\theta_1,\theta_2)) - \frac{a_2^2(\theta_1,\theta_2)}{2\theta_2}\right]$$
$$= V_1(\underline{\theta}) + V_2(\underline{\theta}) + \int_{\underline{\theta}}^{\bar{\theta}} \left(\int_{\underline{\theta}}^{\theta_1} \frac{E_{\theta_2}[a_1^2(s,\theta_2)]}{2s^2} ds \right) f(\theta_1) d\theta_1 + \int_{\underline{\theta}}^{\bar{\theta}} \left(\int_{\underline{\theta}}^{\theta_2} \frac{E_{\theta_1}[a_2^2(\theta_1,s)]}{2s^2} ds \right) f(\theta_2) d\theta_2$$

This is the necessary and sufficient condition for budget-balance and Bayesian incentive compatibility

Substituting for a_1^{FB} and a_2^{FB}, in the previous equation, we get the necessary and sufficient condition for budget-balance, Bayesian incentive compatibility and first-best agreement:

$$V_1(\underline{\theta}) + V_2(\underline{\theta}) = \underline{\theta} \tag{22}$$

Step 2: The next step consists in adding the interim individual rationality constraints.

$$E_{\theta_j}[V_i(y|\theta_i)] \geq V_i^N(\theta_i) = \frac{b_i^2}{2}\theta_i + b_1 b_2 E[\theta] \quad \text{for } i = 1,2 \tag{23}$$

Note that for a mechanism that implements truthfully the first-best abatement level a_i^{FB} the type $\underline{\theta}$ is the critical type since, at this first-best and with incentive

[12] Following Börgers and Norman (2009), if types are independent, for every ex-ante budget-balanced mechanism, there exists an ex-post budget-balanced mechanism such that the allocation rule is unchanged and the interim expected payments are unchanged for all agents. Particularly, for this proof, I use corollary 1 of Börgers and Norman (2009).

compatibility (see Lemma 1),

$$\dot{V}_i^{FB}(\theta_i) - \dot{V}_i^N(\theta_i) = \frac{1-b_i^2}{2} > 0$$

Which implies that $V_i^{FB}(\theta_i) - V_i^N(\theta_i)$ is increasing in θ_i.[13]

Hence, a necessary and sufficient condition for the participation constraint (23) to hold everywhere is that it holds at $\underline{\theta}$, the lowest type. Summing the two individual rationality constraints (23) at the lowest type yields:

$$V_1(\underline{\theta}) + V_2(\underline{\theta}) \geq \frac{b_1^2 + b_2^2}{2}\underline{\theta} + 2b_1b_2E[\theta] \tag{24}$$

Combining (22) and (24), we get the necessary and sufficient condition for the existence of an ex-post efficient (first-best agreement) mechanism that is individually rational, Bayesian incentive compatible and budget-balanced:

$$\underline{\theta} = V_1(\underline{\theta}) + V_2(\underline{\theta}) \geq \frac{1}{2}\underline{\theta} + b_1b_2\bar{\theta}$$

This amounts to check:

$$(b_1+b_2)^2\underline{\theta} \geq \frac{(b_1+b_2)^2}{2}\underline{\theta} + b_1b_2\bar{\theta}$$

Which is equivalent to condition (11) in Proposition 2 because $b_1 + b_2 = 1$. □

Appendix B: One country allows for certification (Proof of lemma 2 and Proposition 3)

Assume that country 1's type θ_1 is publicly known and only country 2 has private information about its type θ_2. A direct revelation mechanism has the following outcome function:

$$y(\theta_1, \hat{\theta}_2) = (a_1(\theta_1, \hat{\theta}_2), a_2(\theta_1, \hat{\theta}_2), t_1(\theta_1, \hat{\theta}_2), t_2(\theta_1, \hat{\theta}_2))$$

where θ_1 is the publicly known type of country 1 (i.e. country 1 cannot lie about its type) and $\hat{\theta}_2$ is the reported type of country 2. The first-best agreement y^{FB} will

[13] Optimal contracting under type-dependent reservation utilities has been extensively analyzed in the literature in the case of a single agent (Lewis and Sappington, 1989 and Jullien, 2000) and in the case of multiple agents (Carrillo, 1998). Here, due to the convexity of the outside option and the fact that the distribution of types is uniform, we are in a simple case in which the critical type is at the bottom of the distribution.

be *implementable* in stage 2 if and only if there exists a direct revelation mechanism $y(\theta_1,.)$ such that (i) the ex-post incentive compatibility and ex-post individual rationality constraints of country 2 are satisfied:

$$\begin{aligned} V_2(y(\theta_1,\theta_2)|\theta_2) &= \max_{\hat{\theta}_2\in[\underline{\theta},\bar{\theta}]} V_2(y(\theta_1,\hat{\theta}_2)|\theta_2) \\ V_2(y(\theta_1,\theta_2)|\theta_2) &\geq V_2^N(\theta_1,\theta_2) = \frac{b_2^2}{2}\theta_2 + b_1b_2\theta_1 \end{aligned}$$

and (ii) the interim individual rationality constraint of country 1 is satisfied:

$$E_{\theta_2}[V_1(y(\theta_1,\theta_2)|\theta_1)] \geq V_1^N(\theta_1) = \frac{b_1^2}{2}\theta_1 + b_1b_2E[\theta] \tag{25}$$

The rest of the proof is similar to the proof of Proposition 2. The first step of the analysis consists in consolidating the budget-balance constraint and the ex-post incentive compatibility constraint for country 2.

Using a similar proof as for Lemma 1, it is easy to show that the direct revelation mechanism $y(\theta_1,.)$ is incentive compatible for country 2 if and only if

$$\dot{V}_2(y(\theta_1,\theta_2)|\theta_2) = \frac{a_2^2(\theta_1,\theta_2)}{2\theta_2^2} \geq 0$$

Integrating this equation yields:

$$V_2(y(\theta_1,\theta_2)|\theta_2) = V_2(y(\theta_1,\underline{\theta})|\underline{\theta}) + \int_{\underline{\theta}}^{\theta_2} \frac{a_2^2(\theta_1,s)}{2s^2}ds \tag{26}$$

Remember that the utility function is given by:

$$V_i(y(\theta_1,\theta_2)|\theta_i) = b_i(a_1(\theta_1,\theta_2) + a_2(\theta_1,\theta_2)) - \frac{a_i^2(\theta_1,\theta_2)}{2\theta_i} + t_i(\theta_1,\theta_2) \tag{27}$$

The ex-ante budget-balance constraint requires that $E_{\theta_2}[t_1(\theta_1,\theta_2)+t_2(\theta_1,\theta_2)] = 0$. Consolidating the budget-balance constraint and equations (26) and (27), and replacing a_1 and a_2 by their first-best levels a_1^{FB} and a_2^{FB} yield:

$$E_{\theta_2}[V_1(y(\theta_1,\theta_2)|\theta_1) + V_2(y(\theta_1,\underline{\theta})|\underline{\theta})] = \left[E[\theta] + \frac{\theta_1}{2} - \frac{\bar{\theta}}{2}\right] \tag{28}$$

This is the necessary and sufficient condition of existence of a first-best agreement that is budget-balanced and ex-post incentive compatible for country 2.

The next step consists in adding the individual rationality (or participation) constraints. We look at the ex-post participation constraint of country 2 and at the interim participation constraint of country 1 for a fixed type θ_1. The interim participation constraint of country 1 is given by :

$$E_{\theta_2}[V_1(y(\theta_1,\theta_2)|\theta_1)] \geq V_1^N(\theta_1) = \frac{b_1^2}{2}\theta_1 + b_1 b_2 E[\theta] \tag{29}$$

Using the same type of argument as in Proposition 2, we can show that $\underline{\theta}$ is the critical type for the ex-post participation constraint of country 2. Hence, a necessary and sufficient condition for the participation constraint of country 2 to hold everywhere is that it holds at $\underline{\theta}$. Therefore, the ex post participation constraint of country 2 is given by:

$$V_2(y(\theta_1,\underline{\theta})|\underline{\theta}) \geq V_2^N(\theta_1,\underline{\theta})$$

At the first-best abatement levels, summing up both participation constraints yields:

$$E_{\theta_2}[V_1(y(\theta_1,\theta_2)|\theta_1) + V_2(y(\theta_1,\underline{\theta})|\underline{\theta})] \geq b_1 b_2(\theta_1 + E[\theta]) + \frac{b_1^2}{2}\theta_1 + \frac{b_2^2}{2}\underline{\theta} \tag{30}$$

Combining (28) and (30), we get the necessary and sufficient condition for incentive compatibility of country 2, individual rationality and budget-balance at the first-best abatement levels:

$$\left[E[\theta] + \frac{\theta_1}{2} - \frac{\bar{\theta}}{2}\right] \geq b_1 b_2(\theta_1 + E[\theta]) + \frac{b_1^2}{2}\theta_1 + \frac{b_2^2}{2}\underline{\theta}$$

This amounts to check:

$$\theta_1 \geq \tilde{\theta}_1 = \frac{b_1}{b_2}(\bar{\theta} - \underline{\theta}) - \frac{b_1^2}{b_2^2}\underline{\theta}$$

which is exactly the condition in Lemma 2.

We get that $\tilde{\theta}_1 = \underline{\theta}$ if the economic environment satisfies:

$$\frac{(\bar{\theta} - \underline{\theta})}{\underline{\theta}} \leq \frac{b_1^2 + b_2^2}{b_1 b_2}$$

And we denote this threshold $\frac{b_1^2+b_2^2}{b_1 b_2} = L^1$. The proof for the situation in which country 2's type is public knowledge and country 1's type is private information is totally symmetric. □

Appendix C: Proof of lemma 4

Before turning to the proof, I define the mechanisms used depending on the actions taken in stage 1. The set of strategies for each country i is $S = \{C,\ NC\}$. Denote by $(s_1, s_2) \in S \times S$ the actions taken in stage 1.

- If $(s_1, s_2) = (C, C)$ (complete information in stage 2), $y(.,.)$ is the mechanism chosen to implement the first-best abatement levels in stage 2 (satisfying budget balance and individual rationality);

- If $(s_1, s_2) = (C, NC)$ (one-sided asymmetric information), $y'(.,.)$ is the mechanism chosen to implement the first-best abatement levels in stage 2 when country 1 allows for certification (satisfying incentive compatibility for country 2, individual rationality and budget balance);

- If $(s_1, s_2) = (NC, C)$ (one-sided asymmetric information), $y''(.,.)$ is the mechanism chosen to implement the first-best abatement levels in stage 2 when country 2 allows for certification (satisfying incentive compatibility for country 1, individual rationality and budget balance);.

Remember that in stage 1, countries do not know their own type. From the best responses derived in the paper, it is clear that at least one country allows for certification at the equilibrium of stage 1. I want to delineate the conditions under which both countries have simultaneously an incentive to allow for certification in stage 1. I thus need to show that it is possible to design a mechanism $y(\theta_1), \theta_2)$ such that at the first-best abatement levels:

- When the strategy of country 1 is $s_1 = C$, the expected utility of country 2 is such that:
$$E_{\theta_1}[V_2(y(\theta_1, \theta_2)|\theta_2)] \geq E[V_2(y'(\theta_1, \theta_2)|\theta_2)] \tag{31}$$
$\forall y'(\theta_1, \theta_2)$ satisfying incentive compatibility for country 2, individual rationality and budget balance.

- When the strategy of country 2 is $s_2 = C$, the expected utility of country 1 is such that:
$$E_{\theta_2}[V_1(y(\theta_1, \theta_2)|\theta_1)] \geq E[V_1(y''(\theta_1, \theta_2)|\theta_1))] \tag{32}$$
$\forall y''(\theta_1, \theta_2)$ satisfying incentive compatibility for country 1, individual rationality and budget balance.

- The budget-balance constraint $t_1(\theta_1, \theta_2) + t_2(\theta_1, \theta_2) = 0$ is satisfied.[14]

Step 1: Rewriting the ex-ante budget-balance constraint

The ex-ante budget-balance requires that:

$$E_{\theta_1\theta_2}[t_1(\theta_1, \theta_2) + t_2(\theta_1, \theta_2)] = 0$$

[14] We can use (as in previous proofs) the ex-ante budget-balance constraint (see Börgers and Norman, 2009).

Using equation (1) and replacing a_1 and a_2 by a_1^N and a_2^N, this constraint can be rewritten as:

$$E_{\theta_1\theta_2}[V_1(y(\theta_1,\theta_2)|\theta_1) + V_2(y(\theta_1,\theta_2)|\theta_2)] = E_{\theta_1\theta_2}\left[\frac{(b_1+b_2)^2}{2}(\theta_1+\theta_2)\right]$$

Or equivalently,

$$\begin{aligned}&E_{\theta_1\theta_2}[V_1(y(\theta_1,\theta_2)|\theta_1) + V_2(y(\theta_1,\theta_2)|\theta_2)]\\ &\quad= E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_1 + b_1b_2\theta_2\right] + E_{\theta_1\theta_2}\left[\frac{b_2^2}{2}\theta_2 + b_1b_2\theta_1\right] + E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_2 + \frac{b_2^2}{2}\theta_1\right]\end{aligned}$$

Step 2: Consider that the strategy of country 1 in stage 1 is $s_1 = C$

Country 2 will also allow for certification if and only if

$$E_{\theta_1\theta_2}[V_2(y(\theta_1,\theta_2)|\theta_2)] \geq E_{\theta_1\theta_2}[V_2(y'(\theta_1,\theta_2)|\theta_2)]$$

with

$$\begin{aligned}E_{\theta_1\theta_2}[V_2(y'(\theta_1,\theta_2)|\theta_2)] &= F(\tilde{\theta}_1)E_{\theta_1\theta_2}\left[\frac{b_2^2}{2}\theta_2 + b_1b_2\theta_1|\theta_1 < \tilde{\theta}_1\right]\\ &+ (1-F(\tilde{\theta}_1))E_{\theta_1\theta_2}\left[\frac{b_2^2}{2}\theta_2 - \frac{b_1^2}{2}\theta_2 + (b_2^2+b_1b_2)\theta_1|\theta_1 \geq \tilde{\theta}_1\right]\\ &+ (1-F(\tilde{\theta}_1))E_{\theta_1\theta_2}[t_2'(\theta_1,\theta_2)|\theta_1 \geq \tilde{\theta}_1]\end{aligned}$$

where $t_2'(\theta_1,\theta_2)$ is a first-best transfer that is incentive compatible and individually rational.

Rearranging terms yields (see equation (14)):

$$\begin{aligned}E_{\theta_1\theta_2}[V_2(y'(\theta_1,\theta_2)|\theta_2)] &= E_{\theta_1\theta_2}\left[\frac{b_2^2}{2}\theta_2 + b_1b_2\theta_1\right]\\ &+ (1-F(\tilde{\theta}_1))E_{\theta_1\theta_2}\left[b_2^2\theta_1 - \frac{b_1^2}{2}\theta_2 + t_2'(\theta_1,\theta_2)|\theta_1 \geq \tilde{\theta}_1]\right]\end{aligned}$$

Step 3: Consider that the strategy of country 2 in stage 1 is $s_2 = C$

Country 1 will also allow for certification if and only if

$$E_{\theta_1\theta_2}[V_1(y(\theta_1,\theta_2)|\theta_1)] \geq E_{\theta_1\theta_2}[V_1(y''(\theta_1,\theta_2)|\theta_1)]$$

with

$$\begin{aligned}E_{\theta_1\theta_2}[V_1(y''(\theta_1,\theta_2)|\theta_1)] &= F(\tilde{\theta}_2)E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_1 + b_1b_2\theta_2|\theta_2 < \tilde{\theta}_2\right]\\ &+ (1-F(\tilde{\theta}_2))E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_1 - \frac{b_2^2}{2}\theta_1 + (b_1^2+b_1b_2)\theta_2|\theta_2 \geq \tilde{\theta}_2\right]\\ &+ (1-F(\tilde{\theta}_2))E_{\theta_1\theta_2}[t_1''(\theta_1,\theta_2)|\theta_2 \geq \tilde{\theta}_2]\end{aligned}$$

where $t''_1(\theta_1,\theta_2)$ is a first-best transfer that is incentive compatible and individually rational. Rearranging terms yields (see equation (14)):

$$E_{\theta_1\theta_2}[V_1(y''(\theta_1,\theta_2)|\theta_1)] = E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_1 + b_1b_2\theta_2\right]$$
$$+ (1-F(\tilde{\theta}_2))E_{\theta_1\theta_2}\left[b_1^2\theta_2 - \frac{b_2^2}{2}\theta_1 + t''_1(\theta_1,\theta_2)|\theta_2 \geq \tilde{\theta}_2]\right]$$

Step 4: Maximum expected transfers

As shown in steps 2 and 3, both $E_{\theta_1\theta_2}[V_1(y''(\theta_1,\theta_2)|\theta_1)]$ and $E_{\theta_1\theta_2}[V_2(y'(\theta_1,\theta_2)|\theta_2)]$ depend on the transfer obtained under the first-best agreement, if implemented, $E_{\theta_1\theta_2}[t''_1(\theta_1,\theta_2)|\theta_2 \geq \tilde{\theta}_2]$ and $E_{\theta_1\theta_2}[t'_2(\theta_1,\theta_2)|\theta_1 \geq \tilde{\theta}_1]$. We want that whatever this transfer, country i always prefers to allow for certification, when the other country also allows for certification. The maximum expected (budget-balanced) transfers that each country can expect when the other country allows for certification is a transfer such that it extracts all the gains from reaching the first-best agreement (the other country is indifferent between the outside option and the first-best agreement):

$$E_{\theta_1\theta_2}[t'_2(\theta_1,\theta_2)|\theta_1 \geq \tilde{\theta}_1] = E_{\theta_1\theta_2}[b_1^2\theta_2 - \frac{b_2^2}{2}\theta_1|\theta_2 \geq \tilde{\theta}_2] \quad (33)$$

$$E_{\theta_1\theta_2}[t''_1(\theta_1,\theta_2)|\theta_2 \geq \tilde{\theta}_2] = E_{\theta_1\theta_2}[b_2^2\theta_1 - \frac{b_1^2}{2}\theta_2|\theta_1 \geq \tilde{\theta}_1] \quad (34)$$

Equations (33) and (34) give the maximum expected transfer that country i may require when country i stays privately informed while country j certifies its type. The maximum expected utility that each country can reach by staying privately informed when the other country allows for certification is then given by:

$$\begin{aligned} E_{\theta_1\theta_2}[V_1(y''(\theta_1,\theta_2)|\theta_1)] &= E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_1 + b_1b_2\theta_2\right] \\ &+ (1-F(\tilde{\theta}_2))E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_2 + \frac{b_2^2}{2}\theta_1|\theta_2 \geq \tilde{\theta}_2\right] \\ E_{\theta_1\theta_2}[V_2(y'(\theta_1,\theta_2)|\theta_2)] &= E_{\theta_1\theta_2}\left[\frac{b_2^2}{2}\theta_2 + b_1b_2\theta_1\right] \\ &+ (1-F(\tilde{\theta}_1))E_{\theta_1\theta_2}\left[\frac{b_2^2}{2}\theta_1 + \frac{b_1^2}{2}\theta_2|\theta_1 \geq \tilde{\theta}_1\right] \end{aligned}$$

Step 5: Consolidating steps 1-4

Remember (step 1) that the ex-ante budget-balance constraint can be written as:

$$\begin{aligned}&E_{\theta_1\theta_2}[V_1(y(\theta_1,\theta_2)|\theta_1)+V_2(y(\theta_1,\theta_2)|\theta_2)]\\&\quad=E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_1+b_1b_2\theta_2\right]+E_{\theta_1\theta_2}\left[\frac{b_2^2}{2}\theta_2+b_1b_2\theta_1\right]+E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_2+\frac{b_2^2}{2}\theta_1\right]\end{aligned}$$

A mechanism $y(\theta_1,\theta_2)$ must satisfy this budget-balance constraint and be such that both countries have an incentive to allow for certification:

$$\begin{aligned}&E_{\theta_1\theta_2}[V_1(y(\theta_1,\theta_2)|\theta_1)+V_2(y(\theta_1,\theta_2)|\theta_2)]\\&\quad\geq E_{\theta_1\theta_2}[V_1(y''(\theta_1,\theta_2)|\theta_1)+V_2(y'(\theta_1,\theta_2)|\theta_2)]\end{aligned}$$

Consolidating these two conditions yields:

$$\begin{aligned}&E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_1+b_1b_2\theta_2\right]+E_{\theta_1\theta_2}\left[\frac{b_2^2}{2}\theta_2+b_1b_2\theta_1\right]+E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_2+\frac{b_2^2}{2}\theta_1\right]\geq E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_1+b_1b_2\theta_2\right]\\&\quad+(1-F(\tilde{\theta}_2))E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_2+\frac{b_2^2}{2}\theta_1\right]+E_{\theta_1\theta_2}\left[\frac{b_2^2}{2}\theta_2+b_1b_2\theta_1\right]\\&\quad+(1-F(\tilde{\theta}_1))E_{\theta_1\theta_2}\left[\frac{b_2^2}{2}\theta_1+\frac{b_1^2}{2}\theta_2\right]\end{aligned}$$

Which is equivalent to:

$$E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_2+\frac{b_2^2}{2}\theta_1\right]\geq(1-F(\tilde{\theta}_2))E_{\theta_1\theta_2}\left[\frac{b_1^2}{2}\theta_2+\frac{b_2^2}{2}\theta_1\right]+(1-F(\tilde{\theta}_1))E_{\theta_1\theta_2}\left[\frac{b_2^2}{2}\theta_1+\frac{b_1^2}{2}\theta_2\right]$$

Or,

$$1\geq(1-F(\tilde{\theta}_2))+(1-F(\tilde{\theta}_1))$$

Replacing $\tilde{\theta}_i$ by its expression (12) (see Lemma 2) yields:

$$\frac{(\bar{\theta}-\underline{\theta})}{\underline{\theta}}\geq\frac{(b_1^2+b_2^2)^2}{b_1b_2(b_1^2+b_2^2-b_1b_2)}=L^2$$

Which is the condition in Lemma 4. □

References

[1] M. Agastya, F. Menezes and K. Sengupta (2007), *Cheap talk, efficiency and egalitarian cost sharing in joint projects*, Games Economic Behavior, 60, 1-19.

[2] S. Baliga and E. Maskin (2003), "Mechanism design for the environment". *In Handbook of Environmental Economics*, Volume 1, ed. K.-G. Maler and J.R. Vincent, 305-324. Amsterdam: Elsevier Science, North-Holland.

[3] R. E. Benedick (1998), *Ozone Diplomacy: New Directions in Safeguarding the Planet*, Harvard University Press.

[4] J.-P. Benoît and J. Dubra (2006), *Information revelation in auctions*, Games and Economic Behavior, 57 (2), 181-205.

[5] T. Börgers and P. Norman (2009), *A note on budget-balance under interim participation constraints: the case of independent types*, Economic Theory, 39 (3), 477-489.

[6] J. Bull and J. Watson (2004), *Evidence Disclosure and Verifiability*, Journal of Economic Theory, 118, 1-31.

[7] J. Bull and J. Watson (2007). *Hard Evidence and Mechanism Design*, Games and Economic Behavior, 58, 75-93.

[8] A. Caparros, J.-C. Pereau and T. Tazdait (2004), *North-South Climate Change Negotiations: A Sequential Game with Asymmetric Information*, Public Choice, 121(3), 455-480.

[9] J.D. Carrillo (1998), *Coordination and Externalities*, Journal of Economic Theory, 78 (1), 103-129.

[10] F. J. Costa and H. A. Moreira (2012), *On the Limits of Cheap Talk for Public Good Provision*. Available at SSRN: http://ssrn.com/abstract=2029331 or http://dx.doi.org/10.2139/ssrn.2029331.

[11] R. Deneckere and S. Severinov (2008), *Mechanism design with partial state verifiability*, Games and Economic Behavior, 64, 487-513.

[12] A. Espinola-Arredondo and F. Munoz-Garcia (2012), *Keeping Negotiations in the Dark: Environmental Agreements under Incomplete Information*, Washington State University Working Papers.

[13] F. Forges and F. Koessler (2005), *Communication equilibria with partially verifiable types*, Journal of Mathematical Economics, 41, 793-811.

[14] J. Hagenbach, F. Koessler and E. Perez-Richet (2014), *Certifiable Pre-play Communication: Full Disclosure*, Econometrica, 82(3), 1093-1131.

[15] C. Helm and G. Wirl (2015), *Climate policies with private information: The case for unilateral action*, Oldenburg Discussion Papers.

[16] M. Hoel and K. Schneider (1997), *Incentives to Participate in an International Environmental Agreement*, Environmental and Resource Economics, 9 (2), 153 - 170.

[17] B. Jullien (2000), *Participation constraints in adverse selection models*, Journal of Economic Theory, 93 (1), 1-47.

[18] C. Kolstad (2005), *Piercing the Veil of Uncertainty in Transboundary Pollution Agreements*, Environmental and Resource Economics, 31 (1), 21-34.

[19] K. A. Konrad and M. Thum (2014), *Climate Policy Negotiations with Incomplete Information*, Economica, 81, 244-256.

[20] J.-J. Laffont and E. Maskin (1979), A differential approach to expected utility maximizing mechanisms, in *Aggregation and Revelation of Preferences*, (J.-J. Laffont, Ed.), 289-308, North-Holland, Amsterdam.

[21] T. Lewis and D. Sappington (1989), *Countervailing incentives in agency problems*, Journal of Economic Theory, 49 (2), 294-313.

[22] G. J. Mailath and A. Postlewaite (1990), *Asymmetric Information Bargaining Problems with Many Agents*, Review of Economic Studies, 57, 351-367.

[23] D. Martimort and W. Sand-Zantman (2015), *A Mechanism Design Approach to Climate Agreements*, Journal of the European Economic Association (forthcoming).

[24] M. McGinty (2007), *International Environmental Agreements among asymmetric nations*, Oxford Economic Papers, 59, 45-62.

[25] R. B. Myerson (1982), *Optimal coordination mechanisms in generalized principal-agent models*, Journal of Mathematical Economics, 10 (1), 67-81.

[26] R. B. Myerson and M. A. Satterthwaite (1983), *Efficient Mechanisms for bilateral trading*, Journal of Economic Theory, 29 (2), 265-281.

[27] T. Palfrey and H. Rosenthal (1991), *Testing for effects of cheap talk in a public goods game with private information*, Games Economic Behavior, 3, 183-221.

[28] R. Rob (1989), *Pollution Claim Settlements under Private Information*, Journal of Economic Theory, 47 (2), 307-333.

[29] U. J. Wagner (2001), *The design of stable international environmental agreements: economic theory and political economy*, Journal of Economic Surveys, 15(3), 377-411.