Working Paper Series

Relational Voluntary Environmental Agreements when Emissions Are Unverifiable

by

Berardino Cesi, Alessio D'Amato

08/2021

SEEDS is an interuniversity research centre. It develops research and higher education projects in the fields of ecological and environmental economics, with a special focus on the role of policy and innovation. Main fields of action are environmental policy, economics of innovation, energy economics and policy, economic evaluation by stated preference techniques, waste management and policy, climate change and development.

The SEEDS Working Paper Series are indexed in RePEc and Google Scholar.
Papers can be downloaded free of charge from the following websites:
http://www.sustainability-seeds.org/.

Enquiries:info@sustainability-seeds.org

SEEDS Working Paper 08/2021
June 2021
By Berardino Cesi, Alessio D'Amato

Relational Voluntary Environmental Agreements when Emissions Are Unverifiable*

Berardino Cesi†, Alessio D'Amato‡

Abstract

Environmental regulation and pollution control may clash against the presence of unverifiable tasks, like source specific emissions. To tackle this issue we reshape a voluntary agreement instrument, already available in the received literature, in a dynamic perspective by means of a relational contracting approach. Setting up a Relational Voluntary Environmental Agreement (REA) helps the regulator to solve the unverifiability issue, and may provide polluting firms with the incentives to stick to environmental requirements. In an N firms symmetric context we show that even if emissions are not contractible across firms, so that enforcement cannot be delegated to a third party, if firms themselves are sufficiently patient, a self-enforcing equilibrium, under which the environmental objective is voluntarily met, exists. Finally, the policy analysis reveals that our REA may be welfare-improving with respect to a Voluntary Environmental Agreement on contractible emissions. This occurs when the enforcement cost savings under a relational agreement are larger than the additional social costs related to free riding.

JEL numbers: D62, H23, Q58. **Keywords:** Relational Contracts, Environmental Policy, Unverifiability, Voluntary Environmental Agreement

*This Version: June 2021.

†University of Rome "Tor Vergata"

‡Corresponding author. University of Rome "Tor Vergata" and SEEDS. Email: damato@economia.uniroma2.it.

1 Introduction

Voluntary approaches to environmental policy are increasingly used worldwide, taking several different forms and covering several environmental realms, including chemicals (e.g. Bi and Khanna, 2012), climate change mitigation (e.g. Salas et al., 2018), agricultural water provision (e.g. Michler and Wu, 2020), among others[1]. The effectiveness of voluntary approaches hinges on appropriate signals for firms to reduce their environmental impact voluntarily (Segerson, 2013; 2018), and this may take place through positive market related incentives as well as through "negative" regulatory based threats. We focus on the latter approach, namely on the role of free riding and the performance of a voluntary agreement in terms of cost effectiveness. Our work extends existing contribution by focusing on a self-enforcing voluntary environmental agreement, that we label as *Relational Environmental Agreement* (REA), linked to the possibility that the voluntary achievement by polluting firms of an exogenous environmental target arises as an equilibrium of a repeated game in contexts where environmental quality is non-verifiable and/or not contractible. This may be the case when environmental quality may be observed (perfectly or imperfectly) but cannot be used as "hard evidence" (e.g. in a court).

Indeed, there are several situations where unverifiability of environmental quality may matter. The enforcement of contracts in the context of global pollution problems may be a crucial issue (e.g. Salas et al., 2018 in relation to REDD contracts). Lacking contractibility may also arise due to weak institutions (as in the case of groundwater contracts in developing countris, e.g. Michler and Wu, 2020, or in relation to conservation agreements, e.g. Gjertsen et al., 2020), as well as in the case of non-point pollution, where the use of an ambient tax, proposed by several papers[2], may be questioned (Sarr et al., 2019), while input

[1] See Segerson (2018) for a detailed review.

[2] The vast literature includes the seminal papers by Segerson (1988) and Xepapadeas (1991). Closer to our analysis, Segerson and Wu (2006) and Suter et al. (2010), deal with the regulation of non-point pollution in a context featuring regulator's threats and voluntary group mechanisms, without focusing, however, on the issue of unverifiability.

taxation may face constraints when unverifiable polluting inputs can replace inputs subject to taxation (Nyborg, 2000). We link to this literature as we model a REA that is at the same time self enforcing and robust to non-verifiability, and show that such environmental agreements may indeed be, at the same time, implementable and social welfare improving as compared to those arising in a static context where verifiability is assumed, due to enforcement costs savings.

In this respect, the closest papers to our analysis are Dawson and Segerson (2008) and McEvoy and Stranlund (2010); they develop similar multiple-firm models (assuming identical firms and a uniform pollutant) to investigate the properties of self-enforcing coalitions. Dawson and Segerson (2008) show that a self-enforcing equilibrium in which firms voluntarily reduce emissions exists but it involves free riding, and thus, cost-inefficiencies. McEvoy and Stranlund (2010) add to their model a costly enforcement component by allowing a third-party to monitor and sanction firms for non-compliance with the voluntary agreement's terms. We extend these works by assuming that firms and the regulator may join a REA where the voluntary agreement arises as an equilibrium of a repeated game, along the lines of the relational contracting literature (Albano, et al. 2017a and 2017b; Cesi et al., 2012; Levin, 2003; Calzolari and Spagnolo, 2009). More specifically, and differently from the static scenario in Dawson and Segerson (2008) and McEvoy and Stranlund (2010), we model a REA extending the analysis to a dynamic open-participation voluntary agreement, where the the latter is self-enforcing and verifiability is not needed across firms. This is expected to be relevant when unverifiability issues are expected to arise in the relationship among firms taking part to the agreement and/or in the delegation of enforcement to a third party (as in McEvoy and Stranlund, 2010). As we will see, however, our REA may also be seen as a welfare improving strategy as compared to a third party-enforced agreement.

Our main results are as follows: first of all, and coherently with the existing literature, we show that free riding by some firms poses a limitation to the possibility that the implementation of a REA is profitable for firms and socially desirable for the regulator. Nonetheless, we prove that, if firms are sufficiently

patient, a self-enforcing equilibrium, under which the environmental objective is voluntarily met, exists. Finally, the policy analysis reveals that our REA may be welfare-improving with respect to a "static" VEA where (costly) enforcement is delegated to a third party, as the one proposed in McEvoy and Stranlund (2010). This is a central contribution of our paper: the mechanism that we suggest is a viable alternative to a static VEA under non-verifiability. Another value added with respect to the existing literature is our explicit focus on the trade-off between the size of the agreement and the discount rate in driving the feasibility of a REA. Our conclusions provide hints that our REA is more likely to be "successful" *ceteris paribus* when regulated sectors feature mature technologies and/or are more stable in terms of entry and exit.

The paper is organized as follows: section 2 presentes the model and the static game, while section 3 presents the repeated game and related main results. Sections 4 and 5 deal with the agreement stability and with welfare analysis, respectively; finally, section 6 concludes.

2 The model

Our model considers an industry composed of N identical firms which produce a specific pollutant. We borrow the functional forms from McEvoy and Stranlund (2010). Strictly concave profits are given by the following function:

$$\pi(e) = \beta + be - (b''/2)e^2$$

where β, b and b'' are strictly positive parameters, with $\pi(0) = \beta$ and e indicates pollution level. It is straightforward to derive the equilibrium emissions and profits level in the "business as usual" (BAU, i.e. unregulated) scenario, given respectively by:

$$e^u = b/b'' \tag{1}$$

and

$$\pi(e^u) = \beta + \frac{b^2}{2b''} \tag{2}$$

If emission taxation is adopted, we assume a regulator sets a per-unit tax t at a level compatible with the achievement of an exogenous total emissions cap $\overline{E}$. Under a unit emissions tax monitoring is costly, but emissions are verifiable. As in McEvoy and Stranlund (2010), we assume a linear enforcement technology, where f is the maximum feasible fine level while α is a measure of the effectiveness of the monitoring technology. We also retain the assumption that enforcement effort and the fine are set in such a way to achieve full compliance. Under these assumptions, per firm enforcement costs guaranteeing perfect compliance are given by $m = \frac{t}{\alpha f}$.[3] As it is reasonable, firm level monitoring costs increase with the tax rate t and decrease with the feasible fine and with the effectiveness of the monitoring technology. Given the assumption of full compliance, firms, choose their emissions level to maximize profits net of tax payment: $\max_e \pi(e^T) = \pi(e) - te$, resulting in the following equilibrium emissions and profits:

$$e^T = (b - t)/b'' \tag{3}$$

$$\pi(e^T) = \beta + \frac{(b-t)^2}{2b''} \tag{4}$$

with $\pi(e^u) > \pi(e^T)$, as it is reasonable. Clearly, in order for the tax to be coherent with the exogenous environmental target, the unit emission tax t should be set according to $\overline{E} = Ne^T = \frac{N(b-t)}{b''}$[4]. In our setting, the emission tax is implemented if voluntary efforts performed by regulated firms are not enough to achieve the environmental target $\overline{E}$. The next section provides details on the modelling strategy for our voluntary environmental policy tools.

[3] The proof mimics closely section 4.1. in McEvoy and Stranlund (2010), and is therefore omitted here. It is based on standard reasoning according to the well established literature on public enforcement of law (see, among others, Polinsky and Shavell, 1999).

[4] We also assume that $b > t$ so that emissions under taxation are strictly positive.

3 The Relational Environmental Agreement (REA)

3.1 The static game

The stage game of our repeated interactions setting mimics the static model presented in McEvoy and Stranlund (2010).

- **Opening Stage**: the regulator decides whether to impose an emission tax**.** Enforcement is assumed to be set at a level compatible with perfect compliance.
- **Agreement Stage**: firms independently choose whether to become part of the agreement; joining the agreement implies that they simultaneously agree on the emissions level e^P in order to match the required standard $\overline{E}$ set by the regulator.
- **Emissions Stage**: members and non-members simultaneously deliver their emissions levels; emissions are observed, and payoffs are collected.

As specified abovem we retain the assumption of verifiability and (costly) observability of emissions in the tax regime. The main difference is instead that we assume emissions to be not verifiable by firms, while in McEvoy and Stranlund (2010), emissions are verifiable in the context of the enforcement delegation of the voluntary agreement so that third-party enforcement is (costly but) indeed possible. Under unverifiability, it is straightforward to show the following:

Proposition 1 *The Nash equilibrium is such that the regulator imposes the unit emission tax, players emit* e^T *and each firm has profits* $\pi\left(e^T\right) = \beta + \frac{(b-t)^2}{2b''}$.

The proof is in line with Proposition 1 in McEvoy and Stranlund (2010) and a simple application of the backwards induction, so that we omit it. In our static three-stage game the regulator is not able to induce participation to the voluntary agreement as an equilbrium, as unverifiability of the emissions within the industry sector induces firms' moral hazard at the last stage. At the

emission stage members choose emissions such that their profits are maximized and if the regulator has imposed the unit emissions tax then firms choose the level $e^T = \frac{b-t}{b''}$. If instead the regulator did not impose the unit emission tax, then the optimal emissions level for each firm is $e_u = b/b''$ even if they agreed to form a voluntary agreement, because their best response is cheating on the emission level in the absence of any verifiable enforcement. At the opening stage the regulator anticipates the firms' behavior and decides to impose the unit emission tax.

3.2 The repeated game

We model an open-participation Voluntary Agreement as a form of relational contracting (Levin, 2003); differently from the the "static" setting in both McEvoy and Stranlund (2010) and Dawson and Segerson (2008), we consider a long-term repeated interaction between the regulator and the firms, which are asked to abate pollutant emissions by means of a Relational Voluntary Environmental Agreement (REA), formalized as a relational contract, offering to avoid emissions taxation on the entire industry if the agreement implies that the emission target $\overline{E}$ is voluntarily met.

We assume here that, contrary to the case of emissions taxation, when firms voluntarily reduce emissions such emissions are unverifiable, so that no delegation to a third party is possible to enforce any voluntary agreement. This amounts to assume that the emission level is not contractible across *private* agents (firms and, possibly, a third enforcement party) and, as a result, relational contracting is the only viable option for a voluntary agreement to be successful. The nature of relational contracting implies, on the other hand, that neither firms nor the regulator are forced to act by specific contractual clauses whose enforcement depends on the verification power of an external court[5]. The

[5] Our setting could be equivalently relevant for cases where emissions are unverifiable in any scenario (voluntary agreement or taxation). In principle, the unverifiability assumption can neither impede nor induce the regulator's choice to opt for a "standard" environmental policy tool (in our setting emissions taxation), and the application of a tax does not require

REA, in order to stand and provide effective incentives for voluntary abatement, must, on the other hand, be self-enforcing for both parties (Levin 2003). To verify the conditions under which our REA produces an effective impact on emissions (i.e. achieving the exogenous environmental target), we now set up the repeated game as an infinite repetition of the three-stage static game previously described. Assume that all involved players have the same discount factor δ. At the end of each period, the players collect their payoff and the regulator observes the industry emissions level E. We define a REA for any history of the game h^t, as i) a tax t, ii) the existence of an agreement by $S \leq N$ firms, and iii) their respective emissions level e. We define the following *grim trigger strategies*, s_R and s_P:

Regulator's strategy, s^R:

- **first stage**: at the first stage of time t, the regulator does not impose the tax scheme if up to $t-1$ the entire industry has delivered the required target $\overline{E}$, otherwise it switches to the tax regime forever after;

Firm's strategy, s^P:

- **Agreement stage**: join the agreement if up to the first period of time t the regulator has not imposed the tax and all other members have agreed on e^P, otherwise exit the agreement for ever after;
- **Emission stage**: once joined the agreement, deliver the required emission level e^P, otherwise, if not joining or if exiting the agreement, choose the level e^{NP} of emissions maximizing own profit for ever after .

Non-members deliver the uncontrolled level of emissions without tax and deliver e^T if an emission tax has been implemented. We keep the assumption

any prove in front of a legal court and it remains one of the discretionary instruments the regulator may choose to implement. Further, and as it is reasonable, it is not possible to question the regulator (or government in general) in case she should decide to apply a soft tax regime, that in our extreme scenario is $t = 0$.

in McEvoy and Stranlund (2010) such that the implementation of the tax is irreversible because such policy is very costly and time-consuming to be designed and approved. This motivates the use of the unforgiving trigger strategy as described above, where the punishment by the regulator to switch to a tax regime lasts forever. This assumption eases the analysis but does not imply a loss of generality. If the regulator could apply the tax only for a short period and then revert to the no tax case, other strategies could be considered, such as *stick and carrots*, whereby the punishment phase lasts for a fixed period of time (let's say T). In this scenario we would obtain different conditions on the discout factor for the equilibrium, which would be more stringent than for the case of a trigger strategy. The intuition is that the shorter the punishment (low T), the higher is the lowest discount factor such that an equilibrium triggering cooperation exists.

Each firm in the REA may deviate from its strategy s_P in the emission stage. Instead of sticking to e^P, each firm can deliver a "non-cooperative" emission level, which corresponds to the unregulated emissions level from (1) and gives the deviating firm unregulated profits from (2), when all other members of the agreement, $(S-1)$, deliver e^P and the non-members, $(N-S)$, deliver e^{NP}.

As prescribed by the strategies s_R and s_P, after a deviation the agreement collapses, the regulator applies the unit emissions tax regime and the profit during the punishment is equal to (4). Clearly, in order for our repeated game analysis to be meaningful, we assume that the equilibrium in the stage game implies non-cooperation; this amounts to assume that with $\pi(e^u) > \pi(e^P) > \pi(e^T)$[6]. Also, a second necessary condition requires that welfare, measured in our setting by profits minus enforcement costs (the latter relevant in the case of taxation) are larger in the cooperative setting than under the punishment (i.e. taxation) phase. This second condition is needed to make sure that the environmental authority indeed prefers the REA, when it is feasible, to imposing a unit emissions tax. Given the number of participant in the REA S, and recalling

[6] In order for this to hold in equilibrium, we assume that $S > \frac{N\sqrt{t}}{\sqrt{2b-t}}$.

the assumption of full compliance under emissions taxation, this amounts to assume:

$$S\pi(e^P) + (N-S)\pi(e^u) \geq N\pi(e^T) - \frac{Nt}{\alpha f}$$

that is, after straightforward algebra:

$$\alpha f \leq \frac{2Sb''}{Nt + S\,(t-2b)}. \tag{5}$$

In other words, enforcement not "too effective" (low feasible fine and/or low monitoring effectiveness)[7].

Following Levin (2003), s_R and s_P rapresent a subgame perfect equilibrium in which REA arises if and only if the following Incentive Compatibility constraint (IC) is satisfied:

$$\frac{1}{1-\delta}\pi(e^P) \geq \pi(e^u) + \frac{\delta}{1-\delta}\pi(e^T) \tag{6}$$

Note that the participation to the REA is made simultaneously by all firms in the agreement stage, while the emission level is decided after observing the number of members (S). Therefore, the firm's deviation may only take place on the agreed level e^P and not on the participation choice. The decision not to participate, in fact, is not an optimal deviation, that is, the firm prefers to deviate only at the emission stage after joining the REA and agreeing on e^P, as deviating by not partecipating to the voluntary agreement would make the agreement unfeasible before any effective choice on emissions is taken, and no short-run gain for the cheating firm would be possible.

In the cooperative phase, given the strategies s_R and s_P, the members maximise their profit, subject to the IC and the emissions target set by the regulator, while non-members free ride and set their emissions level e^{NP} without constrains, so that $e^{NP} = e_u = \frac{b}{b''}$ yielding the profit specified by (2).

[7] In order for condition (5) to be possible, we must assume, as a necessary condition, that $\frac{2S}{N+S}b < t < b$, where the latter inequality is required to have positive emission levels in all scenarios.

As it is common in this kind of repeated games, multiple equilibria arise, so that there may exist several combinations of e and t supporting a *grim trigger strategy* as an equilibrium. The following proposition shows the existence of a self-enforcing REA and the related conditions.

Proposition 2 *Let:*

$$\overline{\delta} = \frac{\pi(e^u) - \pi(e^P)}{\pi(e^u) - \pi(e^T)} = \frac{N^2 t}{S^2(2b - t)} \tag{7}$$

Focusing on solutions such that $e^P > 0$, for any $\delta \geq \overline{\delta}$, s^P and s^R characterise a self-enforcing REA in which members deliver $e^P = \frac{Sb - Nt}{Sb''}$ and no tax is applied. The equilibrium profit is $\pi(e^P) = \beta + \frac{(Sb)^2 - (Nt)^2}{2S^2 b''}$.

Proposition 2 states that when the regulator and firms care enough about future payoffs, then firms have the incentive to agree on a REA that is coherent with emission requirement and that, in turn, avoids the application of an emission tax. The condition (7) gives the lowest possible level of the discount factor such that the REA arises as an equilibrium. Note that $\frac{\partial \overline{\delta}}{\partial t} > 0$, implying that a higher tax rate reduces the willingness of the firms to join a REA and deliver the required emissions. Indeed, an increase in the tax rate decreases $\pi_p(.)$ more than $\pi_t(.)$, reducing the incentives to join the agreement[8].

This analysis takes, however, the agreement size S as given. A possible alternative interpretation of result in Proposition 2 may be derived by taking δ as given, and identifying the minimum agreement size S such that the IC constraint holds. Note that the members' profit is increasing in the size ($\frac{\partial \pi(e^P)}{\partial S} > 0$), and the bigger the size, the lower the abatment for any member to match the target $\overline{E}$. Hence, since $\pi(e^T)$ and $\pi(e^u)$ do not depend on S, it is straightforward to show that the net incentive to participate and stick to the REA is increasing in

[8] Notice indeed that $\frac{d\pi(e^P)}{dt} < 0$ and $\frac{d\pi(e^T)}{dt} < 0$; also, $\left|\frac{d\pi(e^P)}{dt}\right| > \left|\frac{d\pi(e^T)}{dt}\right|$ when $t > S^2 \frac{b}{N^2 + S^2}$, which is always the case, as $t > \frac{2S}{N+S} b > S^2 \frac{b}{N^2+S^2}$. As a consequence, $\frac{d\pi(e^P)}{dt} < \frac{d\pi(e^T)}{dt}$.

S. If we label the lowest value of S satisfying (6) as $S_{\min}$, we get[9]:

$$S_{\min} = \frac{N\sqrt{t}}{\sqrt{\delta(2b-t)}} \tag{8}$$

$S_{\min}$ is the smallest profitable number of members, that we can define as the critical size of the REA for a given δ. When $S < S_{\min}$, members have incentive to deviate and deviation implies the collapse of the REA. The regulator, having complete information over the profit, can perfectly predict such profitability condition on the size of the agreement, so that she would prefer to implement a "standard" emission taxation regime if any agreement with size $S < S_{\min}$ is observed.

4 Agreement stability

We now check whether our profitable REA, characterized by $S \geq S_{\min}$ (or alternatively by $\delta \leq \overline{\delta}$) is stable. More specifically, we check whether our self-enforcing size satisfies internal and external stability conditions. The external stability entails that no non-members have incentive to enter the agreement while internal stability entails that no member has incentive to leave. This concept of stability has been adopted for the environmental agreements in a static setting by Barrett (1994) and follows McEvoy and Stranlund (2010) and Dawson and Segerson (2008). However, hereby we extend the definition of stability to a dynamic scenario. We integrate this definition of stability with the self-enforcing nature of the relational agreement where no deviation exists *on* the cooperative equilibrium path. The definition of both internal and external stability as stated in the current literature does not check for the ex post incentive to effectively deliver the agreed emission once the agreement is formed and the fiscal regime avoided. In a dynamic scenario the self-enforcing property of the agreed emission, in terms of no deviation (from its level) *on* the equilibrium path, is not ensured by this static definition of stability. Once agreed on the

[9] Notice that $S = S_{\min}$ is feasible in our setting, as $S_{\min} > \frac{N\sqrt{t}}{\sqrt{2b-t}}$. In order for the value of $S_{\min}$ to be meaningful, we additionally assume that $\delta \geq \frac{t}{2b-t}$.

cooperative emission level, in fact, each partecipant may have a one shot profitable deviation (OSPD) from the cooperative startegy entailing a reduction in its own emission (e^u in our specific case) without inducing the collapse of the agreement. An agreement size S simply satisfying the static definition of internal and external stability only implies that no member prefers to be outside and non-members deciding not to enter, regardless of whether at the emission stage the agreed emission level is effectively delivered. The definition of stability necessary in a dynamic model combine both concepts of self-enforcing (no deviation *on* the equilibrium path) and internal and external stability as defined in the static scenario. This version of stability is a necessary condition for a stable and profitable REA to exist and works through the self-enforcing value of S defining a stable REA. Let's first define internal and external stability as it follows, for any S:

- *internal stability*: $\pi^P(S) \geq \pi^{NP}(S-1)$, i.e. no member has incentive to leave the agreement if the profit being outside an agreement of size $S-1$ is lower than the profit obtained by being one of the S members;
- *external stability*: $\pi^{NP}(S) \geq \pi^P(S+1)$, each non-member has no incentive to enter the agreement if the profits obtained by being a member in an agreement with size $S+1$ is lower than the profit obtained by remaining outside an agreement of size S.

The following Proposition gives the main result:

Proposition 3 *Strategy s^P characterizes a self-enforcing stable REA when $S = S_{\min}$.*

Proposition 3 shows that $S_{\min}$ is the only stable self-enforcing agreement size (given δ). Any other size does neither satisfy stability nor induce the delivery of the agreed emissions level.

The analysis of the IC constraint, defining how $S_{\min}$ changes with δ in self-enforcing equilibria, offers further insights. We have shown that the agreement

is stable for the smallest size such that the IC constraint is binding. From (8), $S_{\min}$ is decreasing in δ, implying that the stable size of the REA and the discount rate are substitutes in determining the effectiveness of the REA. The intuition is straightforward: when firms give high value to future cooperation or equivalently are concerned about the everlasting application of an emission tax regime, then the agreement does not need a big size to survive and deliver environmental quality targets.

5 Welfare Analysis

We start by comparing the number of participants in stable (and in our case self-enforcing) voluntary agreements from the relevant literature. Labeling $S_{ds} = N\sqrt{t\,(2b-t)}$ the smallest stable agreement in Dawson and Segerson (2008) and comparing it with $S_{\min}$ we get that $S_{\min} < S_{ds}$, given δ.[10]

This result, coupled with results from McEvoy and Stranlund (2010), implies that the equilibrium number of participants to a self-enforcing and stable REA under unverifiability is lower than the one in a static context with verifiable emissions (we label the corresponding level as S_M). In other words, and as it is reasonable, free riding exists even under unverifiability, and $S_{\min} < S_M$. This implies, as a straightforward conclusion, that emissions under the REA for each participating firm are lower than those under a "static" VEA.

Turning to welfare analysis, we start by computing the industry profit under the REA. Due to the exogeneity of the emissions target, this is the same as welfare, which in our setting is intended to measure cost effectiveness of the REA. Clearly, welfare is calculated, given δ, with $S = S_{\min}$, as this is the only

[10] Recall that for $S_{\min} < N$ we need $\delta > \frac{t}{2b-t}$. Also, in order for $S_{ds} < N$ we need $\sqrt{t\,(2b-t)} < 1$.Note that $S_{\min} - S_{ds} = N\frac{-\sqrt{2b\delta-t\delta}\sqrt{2bt-t^2}+\sqrt{t}}{\sqrt{2b\delta-t\delta}}$.To derive a sign for the above difference, notice that the numerator is decreasing in δ (Indeed, $\frac{d\left(-\sqrt{2b\delta-t\delta}\sqrt{2bt-t^2}+\sqrt{t}\right)}{d\delta} = \frac{1}{2\sqrt{2b\delta-t\delta}}(t-2b)\sqrt{-t\,(t-2b)} < 0$); we conclude that the size differential is negative if $\delta > \frac{1}{(t-2b)^2}$ which is always the case under our assumptions to guarantee that $S_{\min} < N$ and to have positive emission levels in all scenarios.

stable and self-enforcing dimension of the REA, namely:

$$W_{REA} = S_{\min}\pi(e^P)\big|_{S=S_{\min}} + (N - S_{\min})\pi(e^u) = \\ = N\left(\beta + \frac{1}{2}\frac{b^2 - \frac{N}{S_{\min}}t^2}{b''}\right)$$

with $\frac{\partial W_{REA}}{\partial \delta} = \frac{\partial W_{REA}}{\partial S}\frac{\partial S_{\min}}{\partial \delta} < 0$.

For the sake of comparison, we can also derive the first best outcome, derived by an equal share of emissions reduction by all the N firms (i.e. a REA with full participation). First best aggregate profits are as follows:

$$W^* = N\left(\beta + \frac{1}{2}\frac{b^2 - t^2}{b''}\right)$$

Clearly, and as it is reasonable, comparing welfare under the "feasible" REA with the first best level we get $W^* > W_{REA}$.

The intuition is that in the first best with full partecipation the loss for the firms that under the REA would free ride is more than offset by the increase in the profits for the members of the REA that under full partecipation benefit by sharing the emission abatment with otherwise free riding firms.

Also, neglecting enforcement imperfections related to an emission tax implementation, tax generates, as it is obvious, a first best outcome. This conclusion is straightforward comparing W^* and equilibrium profits under an emission tax, given by (4). This would apply also in the static context of McEvoy and Stranlund (2010) assuming perfect (i.e. costless) enforcement of emission taxation.

Conclusions are less straightforward if we look at the comparison of our results under the REA with those of a "static" VEA based on costly third party enforcement. We retain from McEvoy and Stranlund (2010) the assumption that the tax revenue is a simple transfer under a social welfare perspective, and we take from the same paper the calculation of the total costs borne by firms taking part to the "static" agreement, which are given by $\frac{Nt}{\alpha_m f_m}$, where α_m is a parameter representing the enforcement effectiveness of the third party VEA enforcer, while f_m is the maximum feasible fine imposed by the same

enforcer appointed by firms taking part to the VEA. Notice that in our case these costs are driven to 0 by the unverifiability assumption: appointing a third party enforcer is useless if emissions are not verifiable in front of a court. If we label welfare arising in McEvoy and Stranlund (2010) as W_M, we can conclude the following:

Proposition 4 *Welfare under a REA is larger than under a members-enforced costly voluntary agreement if the enforcement costs of the latter are sufficiently large and/or if the discount factor is sufficiently small.*

Proof. Assume that enforcement is paid by firms taking part to the VEA in the static context modelled by McEvoy and Stranlund (2010). Also, label emissions arising under the static VEA as $e_M^P = \frac{S_M b - Nt}{S_M b''}$ from Proposition 1; we can rewrite the corresponding welfare as follows:

$$W_M = S_M \pi(e_M^P)\big|_{S=S_M} + (N - S_M)\pi(e^u) - \frac{Nt}{\alpha_m f_m}$$

where the last term is the total payment by firms due to enforcement, as a function of the expected fine $\alpha_m f_m$ for non compliance under third party enforcement, the tax rate t and the total number of firms N (McEvoy and Stranlund, 2010, equation 12). Welfare under the REA is:

$$W_{REA} = S_{\min} \pi(e^P)\big|_{S=S_{\min}} + (N - S_{\min})\pi(e^u)$$

As a result, after simple manipulation, we get that $W_{REA} - W_M > 0$ if

$$\frac{1}{2}\frac{N^2}{S_M S_{\min}}\frac{t^2}{b''}(S_{\min} - S_M) + \frac{Nt}{\alpha_m f_m} > 0,$$

which may be rewritten as:

$$\alpha_m f_m < \left(\frac{S_M S_{\min}}{S_M - S_{\min}}\right)\frac{2b''}{Nt} \tag{9}$$

The proof is concluded by noting that the RHS of (9) is increasing in $S_{\min}$ and, therefore, decreasing in δ. ■

To grasp the intuition for this result, we can rewrite the welfare differential as follows:

$$W_{REA} - W_M = S_{\min}\pi(e^P)\big|_{S=S_{\min}} - S_M \pi(e^P)\big|_{S=S_M} + (S_M - S_{\min})\,\pi(e^u) + \frac{Nt}{\alpha_m f_m}$$

Three impacts of unverifiability can be identified:

1. the REA features a smaller number of participants and, as a result, lower profits for each of them due to the need to reduce emissions more than under the "static" voluntary agreement;

2. the non participants are more numerous in the REA, so that free riding profits are the same for each non member of the agreement, but larger on aggregate under the REA;

3. the REA implies savings in enforcement costs, due to the self-enforcement constraint driven by non-verifiability, as compared to the members-financed voluntary agreement.

Overall, the first two impacts imply a welfare differential against the REA which is larger the larger is the difference in the number of members of the agreement (i.e. the larger is the discount factor and the lower is $S_{\min}$). The third impact, which is stronger the larger are enforcement costs under a static VEA, favours, instead, the REA. Clearly, if no enforcement costs are present, as in Dawson and Segerson, then the members-enforced voluntary agreement always outperforms the REA. This is not, however, the case when unverifiability prevents firms to contract the agreement explicitly.

6 Conclusions

We contribute to the theoretical discussion emerged on the efficiency of voluntary environmental agreements as an alternative to standard environmental policy tools. Although policy makers are increasingly relying on these instruments, the literature is still cautious in assessing their effectiveness. In this paper we deal with the emissions control of a specific pollutant produced by an economic sector, when the exact level of emissions is perfectly observable by the regulator, but it is prohibitively expensive to be verified in front of a legal court by regulated firms. This implies that the enforcement of a voluntary agreement

cannot be delegated formally to a third party, and the only option for voluntary emissions reduction is a long-term Relational Voluntary Environmental Agreement (REA) among the firms which are part of the polluting sector. The REA is structured to provide a benefit for the regulator by ensuring the environmental target is met, and for firms by saving them from a possibly more costly standard policy measure. We first investigate whether a profitable REA can form in our context even when firms have incentive to free ride. We start by proving that cooperation is unfeasible in a static setting because firms cannot be encouraged to profitably cooperate. This differentiates our paper from McEvoy and Stranlund (2010), where the cooperation is possible even in a static setting due to the assumption of verifiability. Allowing for a long-term relationship threatens the firms by the expected profit decrease they may face if emissions taxation is implemented. We define the players' strategies in a dynamic setting, and we show that, if firms are sufficiently patient, there exists a self-enforcing REA such that member firms voluntarily reduce their emissions level, non-members free ride and no tax is applied. This equilibrium is profitable for members which increase their payoff with respect to the situation where emissions taxation is implemented, and have no incentive to deviate; it is even more profitable for non-members that avoid the tax and the costs faced by members. We also derive the minimum number of participants required for a REA to be profitable. We then explore the agreement stability conditions. Profitability is a necessary but not sufficient condition for stability; therefore, we observe which participation levels are internally and externally stable, in the sense that members have no incentive to leave and non-members have no incentive to join. We demonstrate that players' strategies characterise a self-enforcing stable equilibrium for the lowest profitable number of participants. This critical size of the agreement depends on the value they attribute to future profits. If member firms are sufficiently patient, their strategy characterises a self-enforcing equilibrium. We conclude with policy implications by examining whether and under which conditions such measure is welfare-improving with respect to a member-financed "static" voluntary agreement, such as that proposed in McEvoy and Stranlund

(2010). We show this to depend on the enforcement costs of the agreement when emissions are verifiable in the agreement across firms, as well as on the strenght of free riding incentives under the REA.

The main limitation of our paper relies on the assumption of perfect observability of emissions. While the relaxation of this assumption is left for future research, we think our results may provide food for thought in relation to all pollution circumstances where the unverifiability is a crucial issue.

References

Albano GL, Cesi B, Iozzi A. (2017a), "Public procurement with unverifiable quality: The case for discriminatory competitive procedure". *Journal of Public Economics*, 145, pp. 14-26

Albano GL, Cesi B, Iozzi A. (2017b), "Teaching an Old Dog a New Trick: Reserve Price and Unverifiable Quality in Repeated Procurement". Available at SSRN: https://ssrn.com/abstract=3057659 or http://dx.doi.org/10.2139/ssrn.3057659

Bi, X., Khanna, M. (2012). Reassessment of the Impact of the EPA's Voluntary 33/50 Program on Toxic Releases. *Land Economics*, 88(2), pp. 341-361

Calzolari G, Spagnolo G. (2009), "Relational Contracts and Competitive Screening", Centre for Economic Policy Research (CEPR), Discussion Paper No. DP7434 (August)

Cesi B, Iozzi A, Valentini E. (2012), "Regulating Unverifiable Quality by Fixed-Price Contracts". *The B.E. Journal of Economic Analysis & Policy*, 12, Article 40

Dawson N, Segerson K. (2008), "Voluntary agreements with industries: participation incentives with industry-wide targets". *Land Economics*, 84 (February), pp. 97–114

Gjertsen, H., Groves, T., Miller, D. A., Niesten, E., Squires, D., Watson, J. (2020). Conservation Agreements: Relational Contracts with Endogenous Monitoring. *Journal of Law, Economics, and Organization*, ewaa006

Levin J. (2003), "Relational Incentive Contracts". *American Economic Review*, 93, pp. 835-857

McEvoy DM, Stranlund JK. (2010), "Costly enforcement of voluntary environmental agreements". *Environmental and Resource Economics*, 47, pp. 45–63

Michler, J. D., & Wu, S. Y. (2020). Governance and contract choice: Theory and evidence from groundwater irrigation markets. *Journal of Economic Behavior & Organization*, 180, 129-147

Nyborg K. (2000), "Voluntary agreements and non-verifiable emissions". *Environmental and Resource Economics*, 17, pp. 125-144

Salas, P. C., Roe, B. E., & Sohngen, B. (2018). Additionality when redd contracts must be self-enforcing. *Environmental and Resource Economics*, 69(1), 195-215

Sarr, H., Bchir, M. A., Cochard, F., & Rozan, A. (2019). Nonpoint source pollution: experiments on the average Pigouvian tax under costly communication. *European Review of Agricultural Economics*, 46(4), 529-550

Segerson K. (2018), "Voluntary Pollution Control under Threat of Regulation". *International Review of Environmental and Resource Economics*: 11, pp 145-192

Segerson K. (1988), "Uncertainty and incentives for nonpoint pollution control". *Journal of Environmental Economics and Management*, 15, pp. 87–98

Segerson K, Wu J. (2006), "Nonpoint source pollution control: inducing first best outcomes through the use of threats". *Journal of Environmental Economics and Management*, 51, pp. 165-184

Segerson K. (2013), "Voluntary Approaches to Environmental Protection and Resource Management". *Annual Review of Resource Economics*, 5, pp. 161-180

Suter JF, Segerson K, Vossler CA, Poe GL. (2010), "Voluntary-threat approaches to reduce ambient water pollution". *American Journal of Agricultural Economics*, Vol. 92, Issue 4 (August), pp. 1195–1213

Xepapadeas A. (1991), "Environmental policy under imperfect information: Incentives and moral hazard". *Journal of Environmental Economics and Management*, 20, pp. 113–126

7 Appendix

Proof of Proposition 2

Each member chooses its level of emissions by solving the following maximization problem:

$$\underset{e^P}{Max}\ S\pi(e^P)$$

Subject to

$$Se^P + (N-S)e^{NP} \leq Ne^T \tag{10}$$

and

$$\frac{1}{1-\delta}\pi(e^P) \geq \pi(e^u) + \frac{\delta}{1-\delta}\pi(e^T) \tag{11}$$

The Lagrangian function is

$$L = S\pi(e^P) + \lambda(Ne^T - Se^P - (N-S)e^u) + \mu(\frac{1}{1-\delta}\pi(e^P) - \pi(e^u) - \frac{\delta}{1-\delta}\pi(e^T))$$

The Participants FOC are:

$$\frac{\partial L}{\partial e^P} = S\pi\prime(e^P) - S\lambda + \mu\frac{1}{1-\delta}\pi\prime(e^P) \leq 0 \tag{12}$$

with $e^P \geq 0$ and $\frac{\partial L}{\partial e^P}e^P = 0$

$$\frac{\partial L}{\partial \lambda} = Ne^T - Se^P - (N-S)e^u \geq 0 \tag{13}$$

with $\lambda \geq 0$ and $\frac{\partial L}{\partial \lambda}\lambda = 0$

$$\frac{\partial L}{\partial \mu} = \frac{1}{1-\delta}\pi(e^P) - \pi(e^u) - \frac{\delta}{1-\delta}\pi(e^T) \geq 0 \tag{14}$$

with $\mu \geq 0$ and $\frac{\partial L}{\partial \mu}\mu = 0$

We ignore the solution such that $\lambda, \mu = 0$. $\frac{\partial L}{\partial e^P} \leq 0$, as in this case we would get either $e^P = 0$, or $e^P = e^u = b/b''$.We can also exclude $\lambda = 0$ and $\mu > 0$, as in this case (12) is in contradiction with $\mu > 0$ and $S > 0$. We focus therefore on cases where $\lambda > 0$.

If $\mu = 0$, from (13) we get $e^P = \frac{Sb-Nt}{Sb''}$;excluding also in this case $e^P = 0$, FOCs imply $[S\pi\prime(e^P) - S\lambda] = 0$, i.e. $\pi\prime(e^P) = \lambda$ that gives $e^P = \frac{b-\lambda}{b''}$. The corresponding value of λ is $\lambda = \frac{Nt}{S}$. Note that λ must be lower than b because emissions cannot be negative, thus $\frac{Nt}{S} < b$; from (14) we have:

$$\delta \geq \frac{\pi(e^u) - \pi(e^P)}{\pi(e^u) - \pi(e^T)} = \frac{N^2 t}{S^2(2b - t)}$$

which is the minimum value of δ such that s^P and s^R characterize an equilibrium of the repeated game.

Similar conclusions are possible when $\mu > 0$. We have

$$\frac{\partial L}{\partial e^P} = S\pi\prime(e^P) + \mu\frac{1}{1-\delta}\pi\prime(e^P) - S\lambda = 0 \tag{15}$$

$$\frac{\partial L}{\partial \mu} = \frac{1}{1-\delta}\pi(e^P) - \pi(e^u) - \frac{\delta}{1-\delta}\pi(e^T) = 0 \tag{16}$$

$$\frac{\partial L}{\partial \lambda} = Ne^T - Se^P - (N-S)e^u = 0 \tag{17}$$

after subsituting for e^T and e^u, (16) and (17) imply:

$$\delta = \frac{N^2 t}{S^2(2b - t)}$$

and

$$e^P = \frac{Sb - Nt}{Sb'} \tag{18}$$

We substitute (18) in (15) and we derive the value of the two multipliers:

$$\lambda = \frac{(S(1-\delta) + \mu)Nt}{S^2(1-\delta)} \ > 0 \text{ if } \mu > 0$$

or

$$\mu = S(1-\delta)\frac{S\lambda - Nt}{Nt} \text{ with } \lambda > \frac{Nt}{S}$$

Proof of Proposition 3

Since the profit of being part of an agreement and delivering e^P, and since static profits in the punisgment phase and in the unregulated scenario do not depend on S, we can conclude that the incentive to cooperate increases in S. The lowest possible cooperative profit, for $S = S_{\min}$, is:

$$\pi^P(.)\big|_{S=S_{\min}} = \frac{1}{2}\frac{2\beta b'' + b^2 + t\delta\,(t-2b)}{b''}$$

We can therefore conclude that if internal and external stability are satisfied at $S = S_{\min}$ and for $\pi^P(S_{\min})$, then by transitivity they are also satisfied for any higher S. Also, recall that IC is satisfied at values of S equal or larger than $S_{\min}$.

Focusing first on internal stability, for any $S > S_{\min}$ the REA is profitable for all participating firms (due to the IC constraint); as a result, any firm that exits from the REA gains in terms of profits, by free riding on other firms' emissions reduction, but does not make the REA unprofitable for the remaining participants. This is true up to a number of participants $S = S_{\min}$. For this size, exiting the agreement leads to the imposition of the emission taxation policy, and the internal stability is satisfied, as $\pi^P(S_{\min}) \geq \pi^{NP}(S_{\min} - 1)$, due to the fact that for $S = S_{\min} - 1$ the REA does not form and the tax is applied, inducing a lower profit for each firm.

Consider external stability. For $S = S_{\min}$ the REA exists and is internally stable. As non-members receive the "business as usual" profit $\pi(e^u)$, and as cooperative profits increase with S, we can show that even for $S \simeq N$, we have:

$$\pi(e^P) \simeq \beta + \frac{b^2 - t^2}{2b''} < \pi(e^u)$$

Thus, the REA is externally stable for any $S \geq S_{\min}$.