Environ Resource Econ
DOI 10.1007/s10640-006-9058-3

ORIGINAL PAPER

The regulatory choice of noncompliance in emissions trading programs

John K. Stranlund

Received: 31 October 2005 / Accepted: 25 October 2006

Abstract This paper addresses the following question: To achieve a fixed aggregate emissions target cost-effectively, should emissions trading programs be designed and implemented to achieve full compliance, or does allowing a certain amount of noncompliance reduce the costs of reaching the emissions target? The total costs of achieving the target consist of aggregate abatement costs, monitoring costs, and the expected costs of collecting penalties from noncompliant firms. Under common assumptions, I show that allowing noncompliance is cost-effective only if violations are enforced with an increasing marginal penalty. However, one can design a policy that induces full compliance with a constant marginal penalty that meets the aggregate emissions target with lower expected costs. This last result does not depend on setting an arbitrarily high constant marginal penalty. In fact, the marginal penalty need not be higher than the equilibrium marginal penalty under the policy with the increasing marginal penalty, and can actually be lower. Finally, tying the marginal penalty directly to the permit price allows the policy objective to be achieved without any knowledge of firms' abatement costs.

Keywords Compliance · Enforcement · Emissions trading · Monitoring · Transferable permits

JEL classification L51 · Q28

1 Introduction

By exploiting the power of a market to allocate pollution control responsibilities, well-designed emissions trading programs promise to achieve environmental quality goals more cheaply than traditional command and control regulations. It is obvious

J. K. Stranlund (✉)
Department of Resource Economics, University of Massachusetts-Amherst, 214 Stockbridge Hall, 80 Campus Center Way, Amherst, MA 01003, USA
e-mail: stranlund@resecon.umass.edu

though that the full potential of emissions trading cannot materialize if these programs are not enforced well. In recognition of this fact, a sizable theoretical literature exists that examines the consequences of noncompliance and the design of enforcement strategies for emissions trading programs. There is, however, a significant omission in this literature—there is no published work that examines what the level of noncompliance should be for emissions trading programs. To fill this gap, this paper addresses the following question: To achieve a fixed aggregate emissions target cost-effectively, should emissions trading programs be designed and implemented to achieve full compliance, or does allowing a certain amount of noncompliance reduce the costs of reaching the emissions target?

Authors of papers in the literature on compliance and enforcement of emissions trading policies often assume that enforcement is not, or cannot be sufficient to induce full compliance (Malik 1990, 2002; Keeler 1991; van Egteren and Weber 1996; Stranlund and Dhanda 1999; Montero 2002). Others restrict their analyses to full-compliance outcomes (Malik 1992; Stranlund and Chavez 2000; Chavez and Stranlund 2003). In practice we find examples of emissions trading programs with significant noncompliance as well as examples with near-perfect compliance. Montero et al. (2002) argue that the development of an emissions trading program for total suspended particulates in Santiago, Chile has been hampered by weak enforcement and significant noncompliance. On the other hand, several EPA emissions trading programs like the SO_2 Allowance Trading and the NO_x Budget Trading programs were clearly designed to achieve very high rates of compliance and have been successful in achieving this goal (US EPA 2004a, b).

The model of this paper assumes that a regulator chooses a supply of emissions permits and monitoring to check individual firms for noncompliance to minimize the expected costs of inducing a fixed aggregate emissions target. Given a penalty schedule for emissions in excess of permit holdings, the supply of permits and the distribution of monitoring determine individual violation levels. The expected costs of an emissions trading program include not only the firms' aggregate abatement costs and the government's monitoring costs, but also the expected costs of sanctioning noncompliant firms. The expected costs of sanctioning violations have been ignored in the literature on enforcing emissions trading policies. Not only is the assumption that penalizing firms is costly a realistic one, it is also an important determinant of the results of this paper.[1]

Unlike much of the literature on optimal law enforcement (see Polinsky and Shavell 2000 for a review), this work is not concerned with choosing optimal penalty "levels," mainly to avoid focusing attention on the common, but not entirely informative result that penalties should be set as high as possible. Instead the analysis focuses on the choice of penalty structure; that is, we will examine the relative merits of employing an increasing marginal penalty or a constant marginal penalty.

[1] The policy objective of minimizing the costs of achieving an arbitrary environmental target has always been an important objective for analysts and policy makers alike. Montgomery's (1972) seminal work on the efficiency of competitive emissions trading takes this approach. The result that competitive emissions trading minimizes the aggregate abatement costs of reaching an aggregate emissions target is perhaps the main justification for proposing and implementing emissions markets. This paper extends the long line of inquiry into the cost-effective design of environmental policies by including the costs of enforcement in the policy objective.

Under common assumptions in the literature on compliance under emissions trading policies—competitive permit markets and risk neutral firms that always hold a positive number of permits—this paper provides several new results that have important implications for designing and enforcing emissions trading policies. With a given increasing marginal penalty schedule, a simple condition involving the relative marginal costs of monitoring and collecting penalties determines whether the cost-effective level of noncompliance is zero or positive. The fundamental tradeoff in this setting is between allowing greater violations to conserve monitoring costs and inducing greater compliance to reduce the expected costs of collecting penalties. However, this tradeoff does not exist when a constant marginal penalty is employed and firms are motivated to hold a positive number of permits. In this case, it is not possible to increase violations to reduce monitoring, because the amount of monitoring necessary to induce the aggregate emissions standard is fixed. This implies that minimizing the expected costs of achieving an aggregate emissions standard requires eliminating the costs of sanctioning noncompliant firms. That is, full compliance by all firms is cost-effective when a constant marginal penalty is employed.

These results suggest that a positive amount of noncompliance is only cost-effective if violations are punished with an increasing marginal penalty. Thus, the regulatory choice of noncompliance rests on a comparison of the costs of a policy with an increasing marginal penalty that allows for some noncompliance and a policy that induces full compliance with a constant marginal penalty. The resolution of this comparison is straightforward: a policy that achieves an aggregate emissions target with an increasing marginal penalty and that allows some noncompliance is more expensive than an alternative policy involving full compliance and a constant marginal penalty.[2] This last result does not depend on setting an arbitrarily high constant marginal penalty. In fact, the marginal penalty need not be higher than the equilibrium marginal penalty under the policy with the increasing marginal penalty, and can actually be lower.

The cost-effectiveness of full compliance may run counter to one's intuition, and some may be surprised by this conclusion. Many of the papers in the literature on compliance and enforcement of emissions trading programs—indeed most of the much larger literature on optimal law enforcement—have focused on imperfect compliance. Some authors assume that enforcement resources are simply insufficient to induce full compliance. For example, Stranlund and Dhanda (1999) examine the choice of enforcement strategy by a budget-constrained enforcer who does not have sufficient resources to induce full compliance. [3] While limited enforcement resources are certainly a factor in many real instances of environmental policy enforcement, the main result of this paper suggests that in designing an emissions trading program to achieve an aggregate emissions target, regulators should allocate sufficient enforcement resources to achieve full compliance.

[2] There is no work in the literature that compares the efficiency properties of alternative penalty schedules for emissions trading policies. Keeler (1991) provides a positive comparison of emissions trading to emissions standards under exogenous enforcement strategies that involve increasing, constant, and decreasing marginal penalties. In contrast, this paper is concerned with deriving endogenous enforcement strategies and the determination of whether marginal penalties should be increasing or constant. Decreasing marginal penalties are not considered in this paper.

[3] Garvie and Keeler (1994) assume this objective in their analysis of enforcing emissions standards, and Macho-Stadler and Perez-Castrillo (2006) assume the same in their analysis of enforcing emissions taxes.

Another common assumption that is used to preclude full compliance outcomes is that penalties are restricted to be no more than some maximum level. For example, Polinsky and Shavell (2000) motivate their review of the literature on the economics of law enforcement with a standard model that assumes that the penalty for a violation is less than the benefit that some in a population receive from a violation. Obviously, with this assumption full compliance is not possible. Although no upper bound is placed on penalties in this paper, the cost-effectiveness of full compliance and a constant marginal penalty does not depend on the freedom to choose an arbitrarily high marginal penalty. All that is required to make sure that full compliance is a regulatory option is that marginal penalties exceed the prevailing price for emissions permits.[4]

The cost-effectiveness of a constant marginal penalty may also be surprising to some, given that such a penalty is not common in the literature on compliance and enforcement of emissions trading programs. Interestingly, constant marginal penalties appear to be much more common for actual and proposed emissions trading programs than in the literature.[5] In fact, I am not aware of any emissions trading program that punishes noncompliance with an increasing marginal penalty.

The rest of the paper proceeds as follows. Since the analysis of this work involves a standard regulatory model in which the government chooses a policy to which the firms react, Sect. 2 first characterizes firms' emissions and violation choices, given an increasing marginal penalty, a monitoring strategy and a supply of permits. Section 3 then characterizes the government choices of monitoring and permit supply (which together determine violation levels), given an increasing marginal penalty schedule. The analysis turns to a constant marginal penalty in Sect. 4, where the cost-effectiveness of a constant marginal penalty and full compliance over any policy involving an increasing marginal penalty and noncompliance is established. Several issues that arise with incomplete information about firms' abatement costs are discussed in Sect. 5. Tying the marginal penalty to the prevailing permit price allows the regulator to meet its environmental target cost-effectively without any information about the firms' abatement costs. However, dealing with incomplete information about abatement costs when the policy goal is to choose the optimal environmental target

[4] Analysts and policymakers alike stress the importance of making sure that marginal penalties exceed the price of permits (US EPA 2003a). For example, noncompliance in the SO_2 Allowance trading program is penalized with a constant marginal penalty that has always been many times higher than going allowance prices. The penalty was set at $2,000 per ton of emissions in excess of allowances in 1990 dollars, while allowance prices have rarely risen above $200 (US EPA 2004a).

[5] See Boemare and Quirion (2002) for examples of penalties in emissions trading programs. There is quite a lot of variation in how actual constant marginal penalties are set. The SO_2 Allowance program employs a fixed (in real terms) financial penalty. Most papers in the literature on enforcing emissions treading programs, including this one, model sanctions as financial penalties. Another variation of a financial penalty is found in the EPA's recent Clear Skies proposal, which called for a unit penalty that is three times the clearing price in the most recent auction of permits (US EPA 2003b). I demonstrate in Sect. 5 of this paper that tying penalties to going permit prices can help maintain compliance when firms' abatement costs are unknown. Many policies employ an offset penalty whereby a firm's excess emissions in one period are deducted from its allocation of permits in the next period. The SO_2 and Clear Skies programs include a one-to-one offset to complement the financial penalties of these programs. The US EPA's Ozone Transport Commission NO_x Budget Program employs a 3-to-1 offset as its primary penalty for noncompliance. Modeling offset penalties requires a dynamic analysis that is beyond the scope of this paper. See Stranlund et al. (2005) for an analysis of the use of offset penalties to maintain compliance in a dynamic emissions trading program with banking provisions.

is not as straightforward. In this case, it may be efficient to design enforcement strategies that provide a safety valve for firms to escape unexpectedly high abatement costs by choosing to be noncompliant. Section 6 concludes.

2 Individual choices under an increasing marginal penalty

The analysis of this paper is based on a standard model of compliance in emissions trading programs.[6] Throughout consider a fixed set of n heterogeneous, risk-neutral firms. A summary of the costs of all the methods firm i can use to reduce its emissions is given by its abatement cost function, $c_i(e_i)$, which is strictly decreasing and convex in its emissions e_i. The firm is allocated l_i^0 emissions permits initially and chooses to hold l_i permits. Each permit confers the legal right to release one unit of emissions. Assume competitive behavior in the permit market so that all trades take place at a constant price p. The analysis throughout is static.[7]

A regulator has perfect knowledge of each firm's permit holding, but cannot observe emissions without a costly audit. Let π_i denote the probability that the regulator is able to make a determination of i's compliance status. Let us assume, like most other analysts, that monitoring produces a measure of emissions that is accurate enough to judge a firm's compliance status without error. The detection probability is common knowledge and the regulator commits itself to it at the outset. If a firm is noncompliant, its emissions exceed the number of permits it holds and its violation is $v_i = e_i - l_i > 0$. If a firm is compliant, $e_i - l_i \leq 0$ and $v_i = 0$. Violations are penalized according to a quadratic penalty function, $f(v_i) = \phi v_i + \gamma v_i^2/2$, where $\phi > 0$ and $\gamma > 0$. When the analysis turns to a constant marginal penalty schedule in Sect. 4, γ will be set to zero.

Assume that the intercept of the marginal penalty schedule, ϕ, is greater than the equilibrium price of permits. This assumption guarantees that marginal penalties always exceed the price of permits, which allows full compliance to be a possible outcome throughout the paper. No upper bound on marginal penalties is imposed, but none of the results of this paper rely on setting arbitrarily high penalties.

Assuming throughout that each firm chooses positive emissions, firm i's objective is

$$\begin{aligned} &\min_{(e_i,l_i)} c_i(e_i) + p(l_i - l_i^0) + \pi_i\Big(\phi(e_i - l_i) + \gamma(e_i - l_i)^2/2\Big) \\ &\text{subject to } e_i - l_i \geq 0,\ l_i \geq 0. \end{aligned} \tag{1}$$

Restricting the firm to $e_i - l_i \geq 0$ follows from the fact that a firm will never have an incentive to be over-compliant.[8] Letting $\mathcal{L}$ denote the Lagrange equation for (1) and

[6] It is important to note that the model of this paper can easily be applied to other tradable property rights programs. Recent papers by Hatcher (2005) and Chavez and Salgado (2005) are direct applications of the literature on compliance and enforcement of emissions trading to individual transferable fishing quotas. Thus, the results of this paper apply in this context as well.

[7] Stranlund et al. (2005) examine dynamic enforcement of emissions trading programs that allow various forms of permit banking and borrowing. They do not address the regulatory choice of noncompliance, choosing instead to focus on designing enforcement strategies that guarantee full compliance.

[8] If $e_i < l_i$, then the firm could reduce its abatement costs by allowing its emissions to increase to l_i without incurring any costs.

λ_i denote the multiplier attached to the constraint $e_i - l_i \geq 0$, the first-order conditions for a solution to (1) are:

$$\mathcal{L}_e = c_i'(e_i) + \pi_i[\phi + \gamma(e_i - l_i)] - \lambda_i = 0; \tag{2}$$

$$\mathcal{L}_l = p - \pi_i[\phi + \gamma(e_i - l_i)] + \lambda_i \geq 0,\ l_i \geq 0,\ \mathcal{L}_l l_i = 0; \tag{3}$$

$$\mathcal{L}_\lambda = -(e_i - l_i) \leq 0,\ \lambda_i \geq 0,\ \lambda_i(e_i - l_i) = 0. \tag{4}$$

Because the constraint, $e_i - l_i \geq 0$, is linear and the firm's objective is strictly convex when the penalty function is strictly convex, these conditions are necessary and sufficient to identify unique optimal choices of emissions, permit demand, and violation level.

Throughout I assume that enforcement is sufficient to induce each firm to hold a positive number of emissions permits. Then, (3) holds with equality and combining Eqs. 2 and 3 yields $c_i'(e_i) + p = 0$, which uniquely determines the firm's choice of emissions. Note that this decision rule is independent of the enforcement strategy the firm faces, and thus holds whether the marginal penalty is increasing or constant. Since each firm chooses its emissions so that its marginal abatement cost is equal to the going permit price, the permit market will equalize firms' marginal abatement costs. Consequently, whatever level of aggregate emissions results in equilibrium, the aggregate abatement costs of reaching that level of emissions are minimized. Furthermore, in equilibrium the permit price is equal to the aggregate marginal abatement cost function at the resulting level of aggregate emissions. These results are contained in the following lemma. Since the results are well known, the lemma is offered without proof. Note that the lemma holds regardless of whether the marginal penalty is increasing or constant.

Lemma 1 *Each firm chooses its emissions so that* $c_i'(e_i) + p = 0$. *Consequently, in equilibrium,* $p = -C'(E)$, *where* E *is aggregate emissions and*

$$C(E) = \min_{\{e_i\}} \sum_{i=1}^{n} c_i(e_i) \quad \text{s.t.} \sum_{i=1}^{n} e_i = E. \tag{5}$$

Lemma 1 is an important result for the analysis of emissions trading programs generally and for this work in particular. That a competitive permit market leads to a distribution of emissions that minimizes the aggregate abatement costs of reaching an aggregate emissions target has always been one of the main reasons for proposing market-based policies to control pollution. For this work, as long as Lemma 1 holds and an aggregate emissions standard is achieved, alternative policies all have the same minimized aggregate abatement costs. This allows the regulatory choice of noncompliance to be focused solely on minimizing the expected enforcement costs of inducing the aggregate standard.

Apart from the assumption of competitive permit trading, two assumptions guarantee that Lemma 1 holds throughout this work.[9] The first is that each firm holds a

[9] The lemma will not hold in the presence of market power or transaction costs. See van Egteren and Weber (1996), Malik (2002), and Chavez and Stranlund (2003) for analyses of compliance and enforcement of emissions trading programs in the presence of market power. Chavez and Stranlund (2004) analyze compliance and enforcement in the presence of transaction costs.

positive number of permits. If a firm holds no permits, $l_i = 0$ and its violation is $v_i = e_i$. Using (2) and (3) it is straightforward to show that a firm that holds no permits chooses its emissions so that $p \geq -c'_i(e_i)$.[10] If this inequality is strict for some firms, then the marginal abatement costs of the firms will not be equalized and aggregate abatement costs will not be minimized. In Sect. 4, I briefly discuss the possibility that a policy may require that some firms hold zero permits, but that possibility appears rather remote. One may also wonder whether a real firm would ever hold zero permits, given that this would send such an obvious signal of noncompliance to the regulator.

The other assumption that is necessary for Lemma 1 to hold is that firms do not have subjective evaluations of their detection probabilities that depend on their emissions choices and permit holdings. Malik (1990) has shown that if a noncompliant firm's subjective probability of detection is $\pi_i(e_i, l_i)$, with $\partial\pi_i/\partial e_i + \partial\pi_i/\partial l_i \neq 0$, then it chooses its emissions so that its marginal abatement cost differs from the permit price. If this is the case, then it is unlikely that the firms' marginal abatement costs will be equal and that aggregate abatement costs will be minimized. This is the reason that the objective detection probabilities, as determined by the government's monitoring efforts, are assumed to be common knowledge in this paper.

Now let us turn to the compliance decision. Under the assumption that each firm holds a positive number of permits, Eq. 3 holds with equality. Therefore, upon substitution of $v_i = e_i - l_i$, (3) becomes $v_i = (p - \pi_i\phi + \lambda_i)/\pi_i\gamma$ and the requirements of (4) are $v_i \geq 0$, $\lambda_i \geq 0$, and $\lambda_i v_i = 0$. These conditions can be simplified somewhat by showing that cost-effective monitoring requires $\pi_i \leq p/\phi$. To see why, suppose that $\pi_i > p/\phi$ instead. In this case, $p < \pi_i\phi$ so that $v_i = (p - \pi_i\phi + \lambda_i)/\pi_i\gamma \geq 0$ clearly requires $\lambda_i > 0$. In turn, $\lambda_i v_i = 0$ implies $v_i = 0$. Thus, a firm's violation is zero when $\pi_i > p/\phi$. However, monitoring of i can be reduced to $\pi_i = p/\phi$ without affecting the firm's decision to be compliant. To demonstrate this, suppose toward a contradiction that $\pi_i = p/\phi$, but $v_i > 0$. Then, since $p = \pi_i\phi$, $v_i = \lambda_i/\pi_i\gamma > 0$, which requires $\lambda_i > 0$. However, $v_i > 0$ and $\lambda_i > 0$ contradict $\lambda_i v_i = 0$. Therefore, a firm's violation is zero when $\pi_i = p/\phi$ as well as when $\pi_i > p/\phi$. However, since a firm's compliance is achieved with minimal monitoring by setting the detection probability so that $\pi_i = p/\phi$, monitoring at a higher level cannot be efficient.

Furthermore, monitoring to satisfy $\pi_i \leq p/\phi$ implies $\lambda_i = 0$. When $\pi_i < p/\phi$, $p > \pi_i\phi$ implies $v_i = (p - \pi_i\phi + \lambda_i)/\pi_i\gamma > 0$. Consequently, $\lambda_i v_i = 0$ implies $\lambda_i = 0$. We have already seen that $v_i = 0$ when $\pi_i = p/\phi$. In this case, $p = \pi_i\phi$ implies $v_i = \lambda_i/\pi_i\gamma = 0$, which requires $\lambda_i = 0$. Since $\lambda_i = 0$ when we restrict a regulator's choice of monitoring of a firm to $\pi_i \leq p/\phi$, the firm's violation is determined by $v_i = (p - \pi_i\phi)/\pi_i\gamma$. For the determination of the regulatory choice of violations in the next two sections, it is convenient to specify the detection probability that is necessary to induce violation v_i by firm i. Invert $v_i = (p - \pi_i\phi)/\pi_i\gamma$ to obtain $\pi_i(v_i) = p/(\phi + \gamma v_i)$. Note that, given a positive permit price, our assumption that $\phi > p$ guarantees $\pi_i(v_i) \in (0, 1)$. Our results about a firm's violation decision are summarized in the following lemma.

Lemma 2 *Given a linearly increasing marginal penalty, provided that $\pi_i \leq p/\phi$ and i holds a positive number of permits, its violation choice is $v_i = (p - \pi_i\phi)/\pi_i\gamma$. To induce a violation v_i, the regulator must monitor i so that it's detection probability is $\pi(v_i) = p/(\phi + \gamma v_i)$.*

[10] Substituting (2) into (3) and allowing for the possibility that firm i holds no permits yields $\mathcal{L}_l = p - c'_i(e_i) \geq 0$.

Note that a firm's violation choice, $v_i = (p - \pi_i\phi)/\pi_i\gamma$, depends only on the permit price and the enforcement variables, not on its abatement costs.[11] Moreover, since the permit price and the parameters of the penalty schedule do not vary across firms, the function $\pi_i(v_i)$ does not vary across firms. Using these results, Stranlund and Dhanda (1999) have argued that a budget-constrained regulator that seeks to minimize the aggregate violations of heterogeneous risk-neutral firms cannot use differences in the firms' abatement costs to target its monitoring effort.[12] This result also plays an important role in the determination of the cost-effective levels of noncompliance in emissions trading programs.

3 The regulatory choice of noncompliance with an increasing marginal penalty

We are now ready to characterize a cost-effective emissions trading policy that is enforced with a linearly increasing marginal penalty. The regulatory objective is to minimize the sum of aggregate abatement costs, aggregate monitoring costs, and the expected costs of collecting penalties from noncompliant firms, while holding aggregate emissions to a pre-specified target $\bar{E}$. The instruments available to the regulator are the detection probabilities $\pi_i \leq p/\phi,\ i = 1, \ldots, n$, and the aggregate supply of permits L. Since supplying more permits than the aggregate emissions target would lead to aggregate emissions that exceed $\bar{E}$, the regulator's choice of permit supply is restricted to $L \leq \bar{E}$.

As long as the regulator's choices of permit supply and monitoring induce aggregate emissions equal to $\bar{E}$, Lemma 1 reveals that the equilibrium permit price will be $\bar{p} = -C'(\bar{E})$. Moreover, Lemma 1 simplifies the regulator's problem, because competitive permit trading minimizes the aggregate abatement costs of holding aggregate emissions to $\bar{E}$. Therefore, the regulator only needs to minimize the expected enforcement costs of achieving the emissions target with its choices of monitoring and permit supply. Analytically, rather than choosing monitoring and permit supply to accomplish this, it is more convenient to choose the individual violation levels v_i^*, $i = 1, ..., n$. Then, using Lemma 2, these violation levels are used to determine the detection probabilities $\pi(v_i^*) = \bar{p}/(\phi + \gamma v_i^*),\ i = 1, \ldots, n$. Moreover, the individual violations determine the aggregate supply of permits as the solution to $L^* + \sum_{i=1}^{n} v_i^* = \bar{E}$; that is, the supply of permits plus aggregate violations must be equal to the aggregate standard. Clearly, if the cost-effective design involves positive violations to minimize the expected costs of enforcement, then the supply of permits is less than the aggregate emissions target; that is, $L^* < \bar{E}$. On the other hand, if the expected costs of enforcement cannot be reduced by allowing noncompliance, $L^* = \bar{E}$ and each firm is fully compliant.

[11] Stranlund and Dhanda (1999) were the first to demonstrate this result. They show that a parametric increase in a firm's marginal abatement cost function leads it to increase its emissions. However, this change also induces the firm to demand the equivalent number of additional permits, leaving the firm's violation unchanged. Thus, a firm's violation choice is independent of its abatement cost function.

[12] This conclusion does not hold if firms face fixed emissions standards, because the marginal productivity of increased enforcement in reducing violations tends to be higher for firms that have higher marginal abatement costs or who face lower emissions standards (Garvie and Keeler 1994). Murphy and Stranlund (2006) use emissions trading laboratory experiments to test and confirm the hypothesis that firms' violations choices are independent of their abatement costs.

Note that the regulator must know $\bar{p}$ to determine the optimal policy, but that this price is determined from $\bar{p} = -C'(\bar{E})$. Therefore, we must assume that the regulator has complete information about the aggregate marginal abatement cost function. This assumption is maintained in this section and the next. In Sect. 5, however, I will show that tying the marginal penalty directly to the prevailing permit price allows the regulator to achieve the emissions target cost-effectively, despite its uncertainty about aggregate marginal abatement costs.

Now turn to expected enforcement costs. Suppose that the cost of monitoring firm i is $\mu\pi$ (v_i), where μ is a positive constant that does not vary across firms. (I will discuss the consequences of relaxing the assumption of identical monitoring cost functions at the end of this section.) Using $\pi(v_i) = \bar{p}/(\phi + \gamma v_i)$, aggregate monitoring costs are

$$M(v_1, \ldots, v_n) = \sum_{i=1}^{n} \mu\pi(v_i) = \sum_{i=1}^{n} \left(\frac{\mu\bar{p}}{\phi + \gamma v_i} \right). \tag{6}$$

It is clear that aggregate monitoring costs are monotonically decreasing in a firm's violation. This is due to the fact that allowing a higher violation by a firm is accomplished by monitoring it less closely.

In the literature on the economics of law enforcement it is usually assumed that penalties are imposed without cost.[13] In this case, however, because monitoring costs are decreasing in the firms' violations, assuming costless sanctions would lead us to conclude that a cost-effective emissions trading policy would involve maximum violations. This literally suggests setting the aggregate supply of permits equal to zero and eliminating the permit market altogether. Then, each firm would face a zero emissions standard and an expected penalty for each unit of pollution it releases.[14] In reality, however, penalizing firms is likely to be costly. Sanctioning costs will certainly include the administrative costs associated with imposing and collecting penalties. These costs could also include the potentially more substantial costs of government investigations to generate enough evidence to convince a court of a firm's liability, as well as the social costs of firms' efforts to challenge or avoid the imposition of penalties.

Let β be the per-dollar cost of collecting penalties from noncompliant firms. Since the expected penalty for firm i is $\pi(v_i)(\phi v_i + \gamma v_i^2/2)$, expected aggregate sanctioning costs are $S(v_1, \ldots, v_n) = \sum_{i=1}^{n} \beta\pi(v_i)(\phi v_i + \gamma v_i^2/2)$. Substituting $\pi_i(v_i) = \bar{p}/(\phi + \gamma v_i)$ yields

$$S(v_1, \ldots, v_n) = \beta\bar{p} \sum_{i=1}^{n} \left(\frac{\phi v_i + \gamma v_i^2/2}{\phi + \gamma v_i} \right). \tag{7}$$

Allowing individual violations to increase produces countervailing effects on expected sanctioning costs. Holding a firm's detection probability constant, it is

[13] However, see Polinsky and Shavell (1992) for an analysis of how sanctioning costs affect optimal law enforcement. I model the expected costs of collecting penalties in the same way as Polinksy and Shavell.

[14] This is similar to a result obtained by Arguedas and Hamoudi (2004). They show that the optimal emissions standard for a single firm is zero when violations to this standard are punished with an increasing marginal penalty and sanctions are costless.

obvious that expected sanctioning costs increase if it chooses a higher violation. However, allowing a firm's violation to increase is accomplished by reducing the detection probability, which implies a decrease in expected sanctioning costs. It is straightforward to demonstrate that the former effect dominates the latter so that expected sanctioning costs are increasing in individual violations. Since aggregate monitoring costs are decreasing in individual violations, the regulatory choice of noncompliance balances reduced monitoring costs against increased expected sanctioning costs.[15]

Let TE denote total expected enforcement costs. Since Lemma 1 guarantees that the permit market will minimize aggregate abatement costs, the cost-effective distribution of violations is the solution to:

$$\begin{aligned} &\min_{(v_i)_{i=1}^n} \mathrm{TE}(v_1,\ldots,v_n) = M(v_1,\ldots,v_n) + S(v_1,\ldots,v_n) \\ &\text{subject to } \; v_i \geq 0, \; i = 1,2,\ldots,n. \end{aligned} \tag{8}$$

Note that no upper bound on individual violations is specified. This is because we have already precluded the possibility that it might be optimal to choose an individual violation level such that a firm holds no permits. Given this assumption, possible solutions to the regulator's problem are characterized in the following proposition.

Proposition 1 *Given a linearly increasing marginal penalty and the regulatory objective of minimizing the sum of the firms' abatement costs and the expected enforcement costs of holding aggregate emissions to an exogenous standard* $\bar{E}$:

(1) *If* $\beta\phi^2 \geq \gamma\mu$, *then* $v_i^* = 0$ *for each* $i = 1, 2,...,n$. *Furthermore,* $L^* = \bar{E}$ *and* $\pi^* = \bar{p}/\phi$ *for each* $i = 1, 2,...,n$.
(2) *If* $\beta\ \phi^2 < \gamma\mu$, *then*

$$v_i^* = v^* = \frac{\left(2\gamma\mu - \beta\phi^2\right)^{1/2} - \beta^{1/2}\phi}{\beta^{1/2}\gamma} > 0, \; i = 1,2,\ldots,n. \tag{9}$$

Furthermore, $L^* = \bar{E} - nv^* < \bar{E}$ and $\pi^* = \pi(v^*) = \bar{p}/(\phi + \gamma v^*)$ for each $i = 1, 2,...,n$.

Proof Again, individual violations need to only minimize expected enforcement costs, because the permit market itself minimizes aggregate abatement costs. Using (6) and (7), the first-order condition for the determination of the cost-effective violation by firm i can be written as

[15] Simple forms for monitoring costs and the expected costs of sanctioning firms are used to ease the analysis and to highlight the essential aspects of the regulator's choice of noncompliance. These cost functions can be generalized substantially without affecting the main results of this paper. All that is required is that monitoring costs are decreasing and expected sanctioning costs are increasing in individual violations, the sum of the two costs are strictly convex, and if any two firms have the same violation, then their marginal monitoring costs are equal as are their marginal expected sanctioning costs.

$$\partial TE/\partial v_i = \frac{\bar{p}\left\{\beta\phi^2 - \mu\gamma + \beta\left(\phi\gamma v_i + (\gamma v_i)^2/2\right)\right\}}{\left(\phi + \gamma v_i^2\right)} \geq 0, \quad \text{and } (\partial TE/\partial v_i)v_i = 0. \quad (10)$$

To demonstrate part (1) of the proposition note that if $\beta\phi^2 \geq \gamma\mu$, then $\partial TE/\partial v_i > 0$ for $v_i > 0$, which implies $v_i^* = 0$. Moreover, observe that the condition $\beta\phi^2 \geq \gamma\mu$ only involves the parameters of the penalty function and the unit sanctioning and monitoring costs, which do not vary across firms. Therefore, if $\beta\phi^2 \geq \gamma\mu$, then full compliance is required of each firm. Since the cost-effective policy does not allow noncompliance, the aggregate supply of permits is equal to the aggregate emissions standard; that is, $L^* = \bar{E}$. Finally, using $\pi(v_i) = \bar{p}/(\phi + \gamma v_i)$, monitoring should generate the detection probability $\pi^* = \bar{p}/\phi$ for each$i = 1, 2, \ldots, n$.

Turning to part (2) of the proposition, if $\beta\,\phi^{\,2} < \gamma\,\mu$, then $\partial TE/\partial v_i < 0$ for $v_i = 0$. This implies that $v_i^* > 0$ and the first-order condition (10) holds with equality. That is, $\beta\phi^{\,2} - \mu\,\gamma + \beta\,(\phi\,\gamma\,v_i + (\gamma\,v_i)^2/2) = 0$. The sole positive root of this quadratic equation is given by (9). Note that this solution involves the parameters of the penalty function, and the unit sanctioning and monitoring costs, none of which vary across the firms. Therefore, individual violations are positive and uniform across firms; that is, $v_i^* = v^* > 0$, $i = 1, 2, ..., n$. Since aggregate violations are $nv^* > 0$, the aggregate supply of permits is $L^* = \bar{E} - nv^* < \bar{E}$. Lastly, since individual violations are the same, the detection probability $\pi(v^*) = \bar{p}/(\phi + \gamma v^*)$ is also the same for all firms. The proof of part (2) of the proposition is completed by confirming that the second order condition for a solution to (1) holds. The second derivative of TE with respect to i's violation evaluated at v_i^* is $\partial^2 TE/\partial v_i^2 = \bar{p}\gamma\left(2\gamma\mu - \beta\phi^2\right)\Big/(\phi + \gamma v_i)^3$, which is positive given $\beta\phi^2 < \gamma\,\mu$. Since all the cross partial derivatives of TE are zero, expected enforcement costs are strictly convex at $v_i^* = v^*$, $i = 1, 2,...,n$. This completes the proof. QED

Proposition 1 reveals that with an increasing marginal penalty function, whether the cost-effective level of noncompliance is zero or positive depends on the costs of collecting penalties, monitoring costs and the parameters of the penalty function; that is on the relationship between $\beta\phi^{\,2}$ and $\gamma\mu$. If marginal monitoring costs, μ, are relatively high, then the solution calls for a certain amount of noncompliance to conserve monitoring costs. If marginal sanctioning costs, β, are high, then the solution calls for more intense monitoring to induce perfect compliance to eliminate the expected costs of collecting penalties.

Clearly, a distinctive feature of cost-effective noncompliance with an increasing marginal penalty is that violations are uniform across the firms. There are two reasons for this result. First, as noted earlier, firms' equilibrium violation choices are independent of their abatement costs. Thus, if they all face the same penalty schedule and are monitored at the same level, they will all choose the same violation. In fact, uniform monitoring is cost-effective as long as the marginal costs of monitoring and applying sanctions are the same for each firm. It is straightforward to demonstrate that if marginal monitoring and sanctioning costs vary across firms and if the cost-effective policy calls for positive violations by all firms, then the firms with higher marginal monitoring costs or lower marginal sanctioning costs will have higher violations. It is also possible that the optimal policy would involve a mix of compliant and noncompliant firms. That firms' violations will not be uniform when the marginal enforcement cost parameters are not uniform suggests a targeted

monitoring strategy in which firms with lower marginal monitoring costs or higher marginal sanctioning costs are monitored more closely. It is important to reiterate, however, that the justification for a targeted strategy cannot be based on differences in firms' costs of controlling their emissions. Rather, the justification must come from differences in the costs of detecting violations or in the costs of applying sanctions. A rigorous analysis of the causes and consequences of heterogeneous costs of monitoring and sanctioning may be a fruitful area for future research.[16]

4 The cost-effectiveness of a constant marginal penalty and full compliance

Having characterized the regulatory choice of noncompliance with an increasing marginal penalty, we now turn to this choice when a constant marginal penalty is employed. Suppose that violations are punished with a constant marginal penalty ϕ. To guarantee that full compliance is an option, continue to assume that $p < \phi$. Also continue to assume that each firm holds a positive number of permits. Under these conditions I first demonstrate that designing an emissions trading policy so that firms are fully compliant is cost-effective. I then go on to demonstrate that any policy that achieves an aggregate emissions target with an increasing marginal penalty and that allows some noncompliance is more expensive than an alternative policy that uses a constant marginal penalty and induces full compliance.

4.1 The cost-effectiveness of full compliance under a constant marginal penalty

Modifying the first-order conditions for a firm's choices of emissions and permit demand, (2–4), to incorporate the constant marginal penalty and the assumption of nonzero permit holdings yields:

$$\mathcal{L}_e = c_i'(e_i) + \pi_i\phi - \lambda_i = 0; \tag{11}$$

$$\mathcal{L}_l = p - \pi_i\phi + \lambda_i = 0; \tag{12}$$

$$\mathcal{L}_\lambda = -(e_i - l_i) \leq 0,\ \lambda_i \geq 0,\ \lambda_i(e_i - l_i) = 0. \tag{13}$$

Combining Eqs. 11 and 12 yields $c_i'(e_i) + p = 0$. Thus, under the assumption that each firm holds a positive number of permits, Lemma 1 of Sect. 2 holds. Therefore, as in the analysis of the previous section, the permit market minimizes aggregate abatement costs so the regulatory choice of noncompliance is simply a matter of minimizing the expected enforcement costs of holding aggregate emissions to the target $\bar{E}$. Moreover, as long as the regulator uses its choices of permit supply and monitoring to guarantee that aggregate emissions are equal to $\bar{E}$, the equilibrium permit price is $\bar{p} = -C'(\bar{E})$.

Equation (12) makes it clear that to induce emissions choices so that $c_i'(e_i) + \bar{p} = 0$ for each i, the detection probability that each firm faces must satisfy

[16] The possibility that the marginal costs of monitoring may differ among firms is related to the idea that the government may be able to detect the violations of some individuals more easily than others. Bebchuk and Kaplow (1993) were the first to examine heterogeneous probabilities of apprehension in the determination of optimal law enforcement. A recent paper by Macho-Stadler and Perez-Castrillo (2005) assume heterogeneous probabilities of apprehension in their study of enforcing emissions taxes.

$\bar{p} \leq \pi_i \phi$. In fact, given that $\bar{p} \leq \pi_i \phi$ for each firm, a monitoring strategy involving $\bar{p} < \pi_i \phi$ for some firm involves higher monitoring costs than are necessary. To see why, imagine an equilibrium in which $\sum_{j \neq i} l_j$ permits are held by all firms other than firm i. If $\bar{p} < \pi_i \phi$, then i chooses its emissions so that $c_i'(e_i) + \bar{p} = 0$, and (12) and (13) imply that its demand for permits is $l_i = e_i$; that is, it is fully compliant. Of course, equilibrium in the permit market requires $l_i = L - \sum_{j \neq i} l_j$. Monitoring of firm i is inefficient because it can be reduced so that $\bar{p} = \pi_i \phi$ without affecting any of the firms' choices. With $\bar{p} = \pi_i \phi$, firm i continues to choose its emissions so that $c_i'(e_i) + \bar{p} = 0$, but it is now indifferent about the number of permits it holds in the half-closed interval $(0, e_i]$. However, the reduction in the detection probability of i does not change any of the decisions of the other firms, in particular they continue to hold the same number of permits. The permit market clears if firm i also continues to hold the same number of permits, $l_i = e_i$.

Since any policy involving $\bar{p} < \pi_i \phi$ for some i cannot be optimal, the detection probability $\pi = \bar{p}/\phi$ is applied to each firm. With monitoring set in this way, aggregate monitoring costs are $M = \sum_{i=1}^{n} \mu \pi_i = n \mu \bar{p}/\phi$. Note that this is a constant. Furthermore, since aggregate violations are $\bar{E} - L \geq 0$, expected sanctions are $\pi \phi (\bar{E} - L) = \bar{p}(\bar{E} - L)$ and expected sanctioning costs are $S = \beta \bar{p}(\bar{E} - L)$. Since aggregate monitoring costs are a constant, it is not possible to change monitoring to reduce the expected costs of enforcement. The only way to reduce these costs is to reduce expected sanctioning costs. Given the monitoring required to reach the emissions target, expected sanctioning costs can be completely eliminated by setting the supply of permits equal to the aggregate emissions target to make sure that aggregate violations are zero. The preceding analysis proves the following proposition.

Proposition 2 *Given a constant marginal penalty $\phi > \bar{p}$, to minimize the sum of the firms' abatement costs and the expected enforcement costs of holding aggregate emissions to an exogenous standard $\bar{E}$, each firm is monitored so that $\pi^* = \bar{p}/\phi$ and the supply of permits is $L^* = \bar{E}$; that is, the cost-effective level of noncompliance is zero.*

In the case of an increasing marginal penalty we identified a tradeoff between light monitoring and positive violations to conserve monitoring costs and more intense monitoring and full compliance to conserve sanctioning costs. This tradeoff does not exist when a constant marginal penalty is employed, because the amount of monitoring necessary to induce the aggregate emissions target is independent of the amount of noncompliance. Since the costs of monitoring are then fixed, the regulator must focus on minimizing expected sanctioning costs with the aggregate supply of permits. By setting the supply of permits equal to the aggregate emissions target so that all firms are compliant, the regulator reduces expected sanctioning costs to zero.

The cost-effectiveness of full compliance under a constant marginal penalty depends on the assumption that all firms hold a positive number of permits. While this assumption is maintained throughout this work, it is worthwhile to briefly consider, at least qualitatively, whether the full compliance equilibrium can be improved upon by a policy that is designed so that a subset of firms hold zero permits, and hence, are fully noncompliant. Start with a full compliance equilibrium and imagine reducing the monitoring of a subset of firms. These firms will sell off all of their permits because the permit price exceeds the expected marginal penalty they face. Moreover, they will increase their emissions, because the expected marginal

penalty is lower. Holding aggregate emissions to the target, therefore, requires that the emissions of the other firms must fall. To accomplish this, the supply of permits is reduced to just cover the emissions of these firms and they are monitored so that they are fully compliant. Since these firms choose lower emissions than under the original equilibrium, the permit price must rise. Note that allowing noncompliance by a subset of firms has two countervailing effects on monitoring costs. Monitoring effort is reduced for the noncompliant firms, but monitoring must increase for the compliant firms because the permit price is higher.

There are two other costs associated with moving from a full compliance outcome. First, aggregate abatement costs will be higher, because individual firms' marginal abatement costs are no longer equal; the marginal abatement costs of firms that hold permits are higher than for those who do not. Second, expected sanctioning costs increase from zero because some of the firms are now noncompliant.

Clearly, if it is possible to improve on the full compliance equilibrium then the reduction in monitoring costs for the noncompliant firms must be larger than the increase in the monitoring costs of those firms that remain compliant, and the net reduction in monitoring costs must outweigh the increase in aggregate abatement costs and the expected costs of sanctioning the noncompliant firms. While this may be possible, it is probably so only under very limited circumstances.

4.2 The cost-effectiveness of a constant marginal penalty and full compliance

Propositions 1 and 2 together suggest that the regulatory choice of noncompliance in emissions trading programs depends to a large degree on whether an increasing or constant marginal penalty is employed. In particular, a positive amount of noncompliance can only be cost-effective if violations are punished with an increasing marginal penalty. Thus, the regulatory choice of noncompliance rests on a comparison of the costs of a policy with an increasing marginal penalty that allows for some noncompliance and a policy that induces full compliance with a constant marginal penalty.

To conduct this comparison consider a policy with a given increasing marginal penalty function, $f'(v_i) = \phi + \gamma\, v_i$, and suppose that with this penalty function it is cost-effective to allow a positive amount of noncompliance. From part (2) of Proposition 1, individual violations, v, and the detection probabilities, $\pi(v) = \bar{p}/(\phi + \gamma v)$, are the same for each firm. Aggregate emissions are equal to the target $\bar{E}$ and the supply of permits is $L = \bar{E} - nv$. Denote this policy as P.

Now consider an alternative policy, denoted P^a, that achieves the emissions target, but with a constant marginal penalty and full compliance. Of course, given that the penalty schedule under policy P is fixed, P^a would trivially dominate P if we were free to choose an arbitrarily high constant marginal penalty. Doing so would not be informative, so let us place the two polices on equal footing by choosing the constant marginal penalty under P^a to be equal to the equilibrium marginal penalty under policy P. That is, letting ϕ^a denote the constant marginal penalty under policy P^a, $\phi^a = \phi + \gamma\, v$. Given ϕ^a, to induce the aggregate emissions target cost-effectively, Proposition 2 requires that the uniform detection probability under policy P^a is $\pi^a = \bar{p}/\phi^a$ and the aggregate supply of permits is $L^a = \bar{E}$. As usual, both policies minimize the aggregate abatement costs of holding emissions to $\bar{E}$, so whether one is cheaper than the other rests on a comparison of their expected costs of enforcement.

In fact, monitoring costs under the two policies are the same. The detection probabilities are $\pi(v) = \bar{p}/(\phi + \gamma v)$ and $\pi^a = \bar{p}/\phi^a$ under P and P^a, respectively. However, since $\phi^a = \phi + \gamma v$, $\pi(v) = \pi^a$. Since monitoring effort, and hence monitoring costs, of the two policies are equal, they differ only in their expected sanctioning costs. Obviously, since policy P involves a certain amount of noncompliance and policy P^a does not, expected sanctioning costs are positive under P and zero under P^a. Therefore, total expected costs under policy P^a are lower than under policy P. Note that this will be true for lower constant marginal penalties as well. As ϕ^a is decreased, monitoring costs increase because monitoring effort must increase to maintain the equality $\pi^a = \bar{p}/\phi^a$. Policy P^a continues to dominate as ϕ^a is reduced as long as the increase in monitoring costs is less than the expected sanctioning costs under policy P. These results are summarized in the final proposition of this paper.

Proposition 3 *Consider an emissions trading policy that achieves an aggregate emissions target with a linearly increasing marginal penalty and allows for a positive amount of noncompliance. There are other policies with constant marginal penalties that do not exceed the equilibrium marginal penalty of this policy that induce full compliance and achieve the emissions target with lower expected costs.*

The policy significance of Propositions 2 and 3 is quite strong—there appears to be little justification for designing an emissions trading policy that allows a positive amount of noncompliance to the policy. Under the assumptions maintained throughout this work, the cost-effective design of an emissions trading program should involve a constant marginal penalty that exceeds the expected equilibrium permit price by as much as is practicable; the supply of permits should be equal to the aggregate emissions target, and monitoring should be sufficient to induce full compliance.

5 Further discussion: uncertain abatement costs, cost-effectiveness, and efficiency

There is a difficulty with designing an emissions trading policy to satisfy an aggregate emissions standard when there is the potential for noncompliance that has not been addressed in this paper, but that is easily remedied. To make sure that the firms' aggregate emissions meet the aggregate standard, the expected marginal penalty must be equal to the aggregate marginal abatement cost function evaluated at the standard. In the case of a constant marginal penalty, this requires $\pi\phi = \bar{p} = -C(\bar{E})$. Clearly, holding the firms' emissions to the standard requires knowledge of the aggregate abatement cost function. Given a fixed marginal penalty, without complete information about the aggregate marginal abatement cost function the correct detection probabilities cannot be determined. Fortunately, this difficulty is easily overcome if the marginal penalty function is constructed to adjust with the equilibrium permit price.[17] Consider tying the marginal penalty to the permit price by letting $\phi = hp > p$, where $h > 1$ is a constant. Equating the expected marginal

[17] This is in line with how penalties were to be determined in the EPA's proposed (but not enacted) Clear Skies Initiative, which called for a unit penalty that is three times the clearing price in the most recent auction of permits (US EPA 2003b).

penalty to the permit price requires $p = \pi\phi = \pi hp$, which implies that the uniform detection probability is the constant $\pi = 1/h$. With monitoring in this way and a supply of permits $L = \bar{E}$, the emissions target will be reached at least cost without any information about firms' abatement costs. This approach works because the marginal penalty adjusts to aggregate marginal abatement costs through the equilibrium permit price.

The ability to achieve an emissions target cost-effectively without knowledge of the aggregate marginal abatement cost function gives emissions trading programs an advantage over emissions standards and taxes when enforcement is costly and imperfect. Of course, an emissions tax will minimize aggregate abatement costs, and enforcement of the tax can be structured to eliminate sanctioning costs. However, to hold aggregate emissions to a fixed standard, the tax and the expected marginal penalty must be equal to the aggregate marginal abatement cost function at the standard, which requires complete information about abatement costs. An emissions trading policy can overcome this problem by tying the marginal penalty to the permit price.

Sandmo (2002) has shown that it is possible to use imperfect enforcement to minimize aggregate abatement costs when firms face emissions standards. If all firms face a constant expected marginal penalty and all are noncompliant, then the expected marginal penalty serves as the emissions price that equates firms' marginal abatement costs, thereby minimizing aggregate abatement costs. There are two disadvantages of this scheme. The first is the same problem of setting the correct emissions tax with incomplete information about abatement costs—to meet the aggregate emissions standard the expected marginal penalty must be set to the aggregate marginal abatement cost function at the desired aggregate standard. The second problem is that this scheme incurs sanctioning costs, because equating marginal abatement costs under emissions standards requires that all firms be noncompliant. Designing an emissions trading policy to induce full compliance eliminates these costs.

While it is straightforward to deal with incomplete information about abatement costs if the policy goal is to achieve an aggregate emissions target cost-effectively, it is not so easy to deal with this lack of information when the policy goal is to design a policy to balance the expected costs and benefits of emission control.

With complete information about abatement costs, a cost-effective policy of full compliance and a constant marginal penalty can easily be incorporated into the choice of the optimal emissions target. Suppose that emissions are uniformly mixed and cause damage that is characterized by an increasing and convex damage function $D(E)$. The first-best environmental target is chosen to equate marginal damage to aggregate marginal abatement costs. With the necessity of enforcing an emissions trading program, however, the efficient environmental target is no longer first-best. Suppose that a constant marginal penalty is employed. Proposition 2 then implies that whatever level of aggregate emissions is chosen, it will be optimal to issue that number of permits and monitor firms to guarantee full compliance. The detection probability is $\pi = p/\phi$, where $p = -C'(E)$. If the per-firm cost of monitoring is $\mu\ \pi$, total monitoring costs are $n\mu\pi = -n\mu C'(E)/\phi$. The efficient level of emissions then minimizes the social cost function $D(E) + C(E) - n\mu\ C'(E)/\phi$. Assuming that this is strictly convex, the efficient level of emissions is the solution

to $D'(E) + C'(E) = n\mu\ C''(E)/\phi$.[18] Clearly, complete information about aggregate abatement costs is required to determine the optimal level of aggregate emissions.

Only Montero (2002) has analyzed the efficient design of an emissions trading program with uncertain abatement costs that incorporates the need for costly enforcement. He does so as part of his reexamination of Weitzman's (1974) comparison of price (emissions tax) and quantity instruments (tradable permits) under uncertainty to analyze the effects of imperfect compliance on instrument choice. However, Montero assumes at the outset of his analysis that full compliance is never optimal, even in a world of certainty. Of course, the results of this paper suggest that this assumption is overly restrictive.

Nevertheless, his results suggest that the expected marginal penalty for noncompliance can stand in for the safety-valve price envisioned by Roberts and Spence (1976); that is, the expected marginal penalty can serve as the price that firms pay to escape the cap imposed by the supply of emissions permits when abatement costs turn out to be higher than expected.[19] For this reason he finds that the potential advantage of an emissions tax over a transferable permit system is significantly reduced by incomplete enforcement.

Despite Montero's assumption that full compliance can never be achieved, his analysis does point to the possibility that the expected marginal penalty can be used as a safety-valve price instead of simply to maintain compliance. If it is optimal to use the expected marginal penalty in this way, two important conclusions would follow. First, tying the marginal penalty to the going permit price, while a simple remedy to the problem of uncertain abatement costs when the policy objective is cost-effectiveness, is probably not efficient in all cases. This approach would never allow firms to be noncompliant, and therefore, would never allow firms to escape the burden of unexpectedly high abatement costs.

Second, if it is optimal to use the expected marginal penalty as a safety-valve price, it is because it is optimal to design a policy that allows noncompliance in some circumstances. This stands in contrast to the recommendations of this paper. I have argued that there appears to be no reason to design an emissions trading policy that allows for noncompliance when the policy objective is to reach an aggregate emissions target cost-effectively. This is true with and without complete information about firms' abatement costs if the noncompliance penalty can be tied directly to prevailing permit prices. Moreover, there is no reason to allow noncompliance when the policy objective is full efficiency and regulators have complete information about aggregate abatement costs. Therefore, the pursuit of efficient emission trading policies under incomplete information about firms' abatement costs may be the fundamental justification for designing policies that allow noncompliance.

[18] Since this implies $D'(E) > -C'(E)$, the efficient level of emissions is greater than the first-best level. Less control than first best is efficient, because the optimal choice of aggregate emissions internalizes the enforcement costs of maintaining this level of emissions.

[19] See Jacoby and Ellerman (2004) for an informative account of the evolution of the safety-valve concept in the context of controlling greenhouse gas emissions, including using noncompliance penalties in this role. With a simulation study of greenhouse gas control, Pizer (2002) provides a welfare analysis of greenhouse gas quantity limits, greenhouse gas taxes and hybrid approaches that use taxes as a safety valve. Among several interesting results, he demonstrates that even sub-optimal hybrid policies produce large efficiency gains over pure quantity controls.

While the possibility of using enforcement to both induce compliance and to provide a safety valve is intriguing, how to do so optimally has not been dealt with adequately. This topic appears to be an important area for future research.

6 Conclusion

This paper has addressed a fundamental environmental policy question: To achieve a fixed aggregate emissions target cost-effectively, should emissions trading programs be designed to achieve full compliance, or does allowing a certain amount of noncompliance reduce the costs of reaching the emissions target? Using a conventional model of emissions trading with common assumptions, I have argued that allowing noncompliance is only cost-effective if violations are punished with an increasing marginal penalty, but that any such policy is more costly than one that induces full compliance with a constant marginal penalty. The results of this work suggest a strong recommendation for designing emissions trading programs to meet a pre-determined environmental standard: the supply of permits should be equal to the aggregate emissions target, violations should be punished with a constant marginal penalty that exceeds the equilibrium permit price by as much as is practicable, and monitoring should be sufficient to induce full compliance. In addition, tying the marginal penalty directly to the permit price allows the regulator to achieve the aggregate emissions standard without any knowledge of the firms' abatement costs.

While the results of this paper are quite strong, they should be accompanied by certain caveats that deserve further attention. Several of these have been mentioned at various places in the paper, like the causes and consequences of monitoring and sanctioning costs that may vary across firms and the design of an enforcement strategy to provide a safety valve when abatement costs are uncertain. Let me add just a few more. Reconsidering the results of this work when permit trading is imperfect because of market power or transaction costs may be a fruitful exercise. Furthermore, like most analysts, we've assumed that monitoring produces a perfect measure of emissions to determine its compliance status. However, in many situations in which emissions trading might be applied, monitoring errors are possible. Moreover, in many situations emissions cannot be measured directly and must be inferred from observable data on firms' operations. These issues call for a more thorough look at the monitoring function of regulatory enforcement, and the ultimate consequences for the design and implementation of emissions trading programs.

Acknowledgements Primary funding for this research was provided by the U. S. EPA – Science to Achieve Results (STAR) Program grant #R829608. Additional support was provided by the Cooperative State Research Extension, Education Service, U. S. Department of Agriculture, Massachusetts Agricultural Experiment Station, and the Department of Resource Economics under Project No. MAS00871. I thank Carlos Chavez, James Murphy and Charles Mason for helpful comments on earlier drafts of this work.

References

Arguedas C, Hamoudi H (2004) Controlling pollution with relaxed regulations. J Regul Econ 26(1):85–104

Bebchuk LA, Kaplow L (1993) Optimal sanctions and differences in individual's likelihood of avoiding detection. Int Rev Law Econ 13:217–224

Boemare C, Quirion P (2002) Implementing greenhouse gas trading in europe: lessons from economic theory and international experiences. Ecol Econ 43:213–230
Chavez CA, Salgado H (2005) Individual transferable quota markets under illegal fishing. Environ Resour Econ 31(3):303–324
Chavez CA, Stranlund JK (2003) Enforcing transferable permit systems in the presence of market power. Environ Resour Econ 25(1):65–78
Chavez CA, Stranlund JK (2004) Enforcing transferable permit systems in the presence of transaction costs. Department of Resource Economics, University of Massachusetts, Working Paper 2004-3
Garvie D, Keeler A (1994) Incomplete enforcement with endogenous regulatory choice. J Public Econ 55:141–162
Hatcher A (2005) Non-compliance and the quota price in an ITQ fishery. J Environ Econ Manage 49(3):427–436
Jacoby HD, Ellerman AD (2004) The safety valve and climate policy. Energy Policy 32:481–491
Keeler A (1991) Noncompliant firms in transferable discharge permit markets: some extensions. J Environ Econ Manage 21:180–189
Macho-Stadler I, Perez-Castrillo D (2006) Optimal enforcement policy and firm's emissions and compliance with environmental taxes. J Environ Econ Manage 51:110–131
Malik AS (2002) Further results on permit markets with market power and cheating. J Environ Econ Manage 44(3):371–390
Malik AS (1992) Enforcement cost and the choice of policy instruments for controlling pollution. Econ Inquiry 30:714–721
Malik AS (1990) Markets for pollution control when firms are noncompliant. J Environ Econ Manage 18:97–106
Montero J-P (2002) Prices versus quantities with incomplete enforcement. J Public Econ 85:435–454
Montero J-P, Sanchez JM, Katz R (2002) A market-based environmental policy experiment in Chile. J Law Econ 45(1):267–287
Montgomery WD (1972) Markets in licenses and efficient pollution control programs. J Econ Theor 5(3):395–418
Murphy JJ, Stranlund JK (2006) A laboratory investigation of compliance behavior under tradable emissions rights: implications for targeted enforcement. J Environ Econ Manage (Forthcoming)
Pizer WA (2002) Combining price and quantity controls to mitigate global climate change. J Public Econ 85:409–434
Polinsky AM, Shavell S (1992) Enforcement costs and the optimal magnitude and probability of fines. J Law Econ 35(1):133–148
Polinsky AM, Shavell S (2000) The economic theory of public enforcement of law. J Econ Literat 38(1):45–76
Roberts MJ, Spence M (1976) Effluent charges and licenses under uncertainty. J Public Econ 5:193–208
Sandmo A (2002) Efficient environmental policy with imperfect compliance. Environ Resour Econ 23(1):85–103
Stranlund JK, Dhanda KK (1999) Endogenous monitoring and enforcement of a transferable emissions permit system. J Environ Econ Manage 38(3):267–282
Stranlund JK, Chavez CA (2000) Effective enforcement of a transferable emissions permit system with a self-reporting requirement. J Environ Econ Manage 18(2):113–131
Stranlund JK, Costello C, Chavez CA (2005) Enforcing emissions trading when emissions permits are bankable. J Regul Econ 28(2):181–204
US Environmental Protection Agency (2004a) Acid rain program 2003 compliance report. US EPA Acid Rain Program, Washington DC
US Environmental Protection Agency (2004b) NO_X Budget trading program: 2003 progress and compliance report. US EPA Office of Air and Radiation, Clean Air Market Programs, Washington DC
US Environmental Protection Agency (2003a) Tools of the trade: a guide to designing and operating a cap and trade program for pollution control. US EPA Office of Air and Radiation, Washington DC
US Environmental Protection Agency (2003b) Section-by-section summary of the clear skies act of 2003. US EPA Office of Air and Radiation, Washington DC
van Egteren H, Weber M (1996) Marketable permits, market power, and cheating. J Environ Econ Manage 30:161–173
Weitzman M (1974) Prices vs. quantities. Rev Econ Stud 41(4):477–491