

CHAPTER 4

ECONOMIC PERSPECTIVES ON ENFORCEMENT POLICY

This chapter examines the insights provided by economic theory for the design of policies to enforce environmental regulations. Section 4.1 briefly reviews the economic literature on the enforcement of rules and regulations. Much of this literature addresses issues that are of limited relevance to the problem of enforcing environmental regulations. Therefore, a model of optimal enforcement is developed in Section 4.2. This model attempts to capture the salient features of the various models presented in the economic literature, without sacrificing simplicity. In addition, the model is specifically tailored to the problem of enforcing environmental regulations. The model is primarily designed to examine the behavior of firms that attempt to maximize their profits exclusively. Therefore the types of firms to which the model developed in this chapter is relevant are those that will not tend to comply fully with CWA regulations unless they believe that compliance will enhance their profits. Thus, the enforcement model developed in this chapter is clearly not relevant to firms that would comply with environmental regulations even if they could boost their profits by not engaging in costly compliance expenditures. Hence, this model of optimal enforcement is presented in order to study how best to handle the subset of firms that will not comply with CWA regulations unless their profits are adversely affected by noncompliance.

The exclusively profit-maximizing subset of firms (for which this model is applicable) weigh the costs of compliance against the probability of apprehension and the fines faced if noncompliance is detected. If the fines and probabilities are too low relative to the costs of compliance, these firms may elect to violate the regulations, in a sense, gambling that their noncompliance will not be detected and punished. The model developed in this chapter elaborates on this basic point, outlining the decision making of this exclusively profit-maximizing subset of firms, the cost and benefit consequences for society of their decisions, and the optimal response of EPA in setting fines and allocating scarce enforcement resources. That is, EPA's enforcement policy consists of two pieces -- fines and other types of penalties that are imposed when an enforcement action is taken, and enforcement activities (e.g., monitoring, testing, record-checking, inspections, etc.). The former has received more attention under the guise of "penalty policy", but the latter, "enforcement strategy", is of equal importance in the overall enforcement framework.

The model of these two prongs of enforcement policy developed in this chapter does not provide a simple formula for calculating the optimal penalties for noncompliance and the precise enforcement strategies in terms of exact conclusions for targetting enforcement resources, it offers a number of insights regarding the design of an optimal enforcement policy. The analysis suggests that the optimal method for targetting scarce enforcement resources depends on four related factors:

- the costliness of enforcement (i.e., how expensive it is to catch and fine violators);
- the economic value of the damages resulting from violations to human health and environmental quality;
- the costs to violators of achieving compliance; and
- the degree to which increased enforcement efforts in a given industry or area increase the perceived probability of detection and penalization.

The analysis suggests that enforcement resources (i.e., seeking penalties and detecting violators) should be focused on noncompliance that causes larger damages to the environment, in areas in which the costs of enforcement are relatively lower, and on firms and areas in which substantial changes in the perceived probability of apprehension for noncompliance result from enforcement activities (although the impact of higher compliance costs on the optimal level of enforcement is ambiguous). The framework also strongly suggests that the penalties for failing to report a violation (e.g., falsifying data) must be set jointly with the penalties associated with simple noncompliance. Otherwise, it is possible for the penalty structure to cause firms to falsify data and to conceal violations. The body of this chapter develops and discusses these conclusions in greater depth.

In Section 4.3, the analysis is extended to consider the problem of noncompliance when self-monitoring/reporting is required. The analysis reveals the importance of properly structuring the penalties for exceeding an effluent limit and for failing to report an effluent limit violation. Finally, Section 4.4 discusses the general implications of economic theory for the design of effective enforcement policies. An appendix to this chapter contains an example demonstrating the calculation of the optimal expected fine for an effluent limit violation.

The discussion in this chapter is fairly technical. For readers who would like to review the major conclusions of the chapter without delving into the details of the analysis, a non-technical summary is presented below. This summary concludes with a review of the implications of the economic model for both enforcement strategy (e.g., targetting of enforcement resources) and penalty policy (i.e., optimal setting of penalties).

4.0 NON-TECHNICAL SUMMARY

A review of the economic literature identified a small body of literature relevant to the economics of enforcing environmental regulations. The emphasis of this literature is on analyzing the behavior of firms (dischargers) that do not fully comply with various types of environmental regulations because their commitment to compliance is too weak in the absence of strong profit-related incentives to comply. However, only limited attention is given in the existing literature to the problem of how to

optimally enforce existing environmental regulations. Therefore, a new model of optimal enforcement of environmental regulations, in particular, effluent limit regulations, is developed for this analysis.

The model developed has two key variables that are controlled by the relevant enforcement authority:

- the penalty (fines or other penalties) for effluent limit violations; and
- the perceived probability, or perceived frequency with which firms believe that they will be caught exceeding their effluent limits and penalized for doing so.

These two variables together constitute an enforcement policy. Neither alone is sufficient since both the size of the penalties levied and the perceived probability that they will be levied are both central to determining the degree of compliance likely to be observed on the parts of firms that require financial incentives to comply with environmental regulations. Thus, the penalty for violations times the perceived probability that violations will be caught and penalized is defined to be the expected penalty, that is, the penalty that a firm believes it will pay for violations. Some of a firm's violations are likely to go undetected and unpunished, whereas others will be detected and punished. This uncertainty is captured by the expected penalty variable, because the expected penalty essentially discounts the penalty for violations by the perceived probability that violations may be detected and punished.

An interesting feature of the model of enforcement developed in this chapter is that an analytical distinction is made between the perceived probability of detection and penalization and the objective probability. The former concept is the one that regulates the behavior of firms since it is the perceived expected fine that helps to determine the degree to which firms will comply with environmental regulations. The objective, or actual probability, on the other hand, is the true probability of being detected and penalized. These could be different depending on the information available to firms concerning past enforcement actions and future enforcement expectations. Indeed, some enforcement actions are undertaken precisely because it is felt that firms will greatly increase their expectations regarding the probability of being caught and fined.

The expected penalty is the key parameter influencing a firm's decision on whether to comply with a regulation, given that the firm is one that falls into the subset of firms that will not comply with environmental regulations without strong profit-related incentives to do so. Although the size of the penalty is important, it alone does not determine compliance or noncompliance. An extreme case demonstrating this is one where a very high fine is set, but no resources are devoted to monitoring discharges and detecting violations. In this case, the probability that firms are caught and penalized is virtually zero. If the firms involved also perceive that the probability is nearly zero, then the size of the fine is of little importance

because it will almost never be levied. Hence, for firms that do not require financial incentives to comply with regulations, the penalties are unimportant because these firms always comply. However, for the subset of firms that do require such incentives to comply, if the probability of detection is perceived to be virtually zero, then the impact of noncompliance on the firms' profits is positive (compliance costs are avoided and no penalties are levied).

Thus, an enforcement authority must not only determine the appropriate penalty to set for violations, but must also determine the appropriate frequency with which dischargers should be monitored and penalized for violations in order to properly affect firms' perceptions of this probability. Although attention to date has focused on the appropriate penalties for violations (penalty policy), equal attention should be given to firms' perceptions regarding the frequency with which firms will be monitored and penalized for noncompliance (enforcement strategy). Together these form a coherent enforcement policy.

The analysis shows that the optimal values of the fine and its perceived probability (i.e., the values of these variables at which the benefits minus the costs of increased enforcement are maximized) depend on four factors:

- the costliness of enforcement (i.e., how expensive it is to catch and fine violators);
- the economic value of the damages resulting from violations to human health and environmental quality;
- the costs to violators of achieving compliance; and
- the degree to which increased enforcement efforts in a given industry or area increase the perceived probability of detection and penalization.

The precise value of the optimal fine, as well as the optimal amount of enforcement activity (which determines the perceived probability of detection and penalization), both depend on these four factors in a fairly complicated way. For instance, it is not generally true that the optimal fine is equal to the sum of the benefits from noncompliance (i.e. the compliance costs avoided), the damages due to noncompliance, and the costs of enforcement.

The analysis reveals that setting the penalty equal to the value of the benefits from noncompliance may do little to deter noncompliance if firms do not believe that violations are not always detected and fined. In these cases, it may be in the discharger's interest to exceed effluent limits despite the attendant penalties, given that the firm requires financial incentives to comply with the regulations.

In terms of targetting the enforcement resources of the Agency, the analysis indicates that resources should be focused (1) on violators that impose relatively high damages, (2) on violators against whom it is relatively

inexpensive to bring enforcement actions, and (3) in those areas in which relatively small enforcement expenditures yield relatively large increases in the perceived probability of detection and penalization (the implications of higher compliance costs per se on the level of enforcement costs, are ambiguous however). More precisely, if there are two or more violators that have similar compliance costs and impose similar enforcement costs on the Agency, then enforcement resources should be targetted at the firm or firms whose violations result in the largest damages. Similarly, if there are two or more violators that have roughly equal compliance costs and impose roughly the same damages, then enforcement resources should be targetted at the firm or firms against whom it is least costly to take action. Finally, if two types of dischargers impose the same damages, have the same compliance costs, and cost the same to monitor and to bring enforcement actions, then enforcement resources should be targetted more closely at the industry whose perceptions of the probability of detection increase more rapidly with the underlying objective probability.

The analysis of self-monitoring/reporting requirements demonstrates that if firms are to have an economic incentive to report violations, the penalty for not reporting an effluent limit violation must generally be far larger than the penalty for the effluent limit violation. Otherwise, it is in the discharger's interest to conceal violations, given that the firm decides not to comply with the regulations, which suggests that the penalties for failing to report violations should be set jointly with the penalty for effluent limit violations.

Implications of the Model for Enforcement Policy

Placing the conclusions of the economic model of optimal enforcement in the context of EPA's enforcement of CWA regulations, several general conclusions emerge. These fall into the following three categories:

- Targetting Enforcement Resources;
- Vigorous Enforcement of Self Monitoring/Reporting Requirements; and
- Refining Penalties for Violations.

The first category concerns enforcement strategy, in the sense that it refers to how Agency resources might be best utilized to achieve maximum compliance and, presumably, the greatest environmental benefits. The second and third categories concern penalty policy, (i.e., how penalties might be adjusted to ensure that future noncompliance is deterred). Each category is discussed below.

Targetting of Monitoring Resources

As discussed in Chapter 2, the enforcement process has three major steps:

- Monitoring compliance and detecting violations;

- Taking action against violators -- seeking penalties, if necessary; and
- Following up on violators to ensure that they undertake the agreed upon efforts to limit future violations.

Our study indicates that the first step in this process may well be the most problematic. In general, the difficulty of monitoring compliance and detecting violations depends on the form of noncompliance. The failure of a facility to regularly submit a discharge monitoring report is not difficult to detect; it simply requires checking the facility's submissions against the relevant schedule. Similarly, determining whether or not a facility has installed specific types of abatement equipment can be accomplished with relative ease. Detecting effluent limit violations, on the other hand, is not as simple because it requires continuous monitoring and analysis of a facility's discharges. Given the difficulty and expense of continuously monitoring discharges, this is typically achieved by means of "grab", or composite, sampling of discharges, which only provide a "snapshot" of a facility's compliance status. Given the large number of dischargers and constraints on the resources available for monitoring discharges, sampling of discharges by federal and state officials is carried out relatively infrequently.

The large share of the burden for monitoring discharges is placed on the dischargers themselves. Dischargers are required to report significant violations and to periodically submit discharge monitoring reports even if they are in compliance. If dischargers complied perfectly with these self-monitoring/reporting requirements, detecting violations would not be a problem. However, because (detected) violations bring the threat of enforcement action, firms may be reluctant to report violations and submit discharge monitoring reports. Or, if they do report violations, there is an incentive for dischargers to understate the extent of their violations. Therefore, to ensure that firms report violations, or that they report them accurately, it is essential for the Agency to routinely monitor and analyze discharges.

The problem, as noted earlier, is that monitoring and analyzing discharges is costly because there are thousands of dischargers to be monitored. Since only a limited amount of resources can be devoted to monitoring efforts, the problem becomes one of determining how frequently different dischargers should be monitored by the Agency or state authorities. The focus of monitoring efforts should clearly be on dischargers (1) that are likely to be noncompliant, and, within this group, on dischargers that are likely to impose relatively large damages due to noncompliance, and (2) against which enforcement action is likely to be relatively inexpensive. The Agency has already gone a long way in this regard by developing the major/minor discharger classification and developing criteria for identifying significant violations. There may be scope for more targetting along the following lines:

(1) Technical Criteria that Correlate With Noncompliance -- It is possible that there are technical aspects of production processes or effluent control that correlate with noncompliance. For example, it could be that firms whose production processes generate different types of effluents at different times may be more likely to be in noncompliance than firms whose processes generate the same level and types of effluent most of the time. If such criteria can be identified, this suggests that the technical characteristics of a discharger's production and treatment process may be one useful criterion for targetting monitoring resources.

(2) Unannounced Inspection Visits -- A recent survey of state enforcement agencies conducted by Resources for the Future (Russell, Harrington, and Vaughan, 1985) indicates that the agencies frequently notified dischargers of upcoming inspection visits; only a small fraction did not do so as a matter of policy. If firms are able to alter the quantity or composition of their waste streams on short notice, the compliance status of a discharger observed during an inspection visit may not present an accurate picture of the discharger's day-to-day compliance status. Dischargers may step up treatment processes during inspection visits and shut down particularly noxious production processes to limit the extent of any violations with permit requirement. On the other hand, OWEPP recommends that firms be notified that an inspection visit will occur within the next six months, but should not be told when the visit precisely the visit will occur. Analytically, this is equivalent to unannounced inspection visits, as recommended here.

(3) Tying Inspection Frequency to Past Behavior -- Currently, inspection frequencies are primarily determined by the classification of a discharger as a major or a minor discharger. The survey of state agencies referred to above indicates that major dischargers are inspected roughly four times a year, while minor dischargers are inspected on the order of once a year. It does not appear that the past behavior of dischargers is routinely incorporated as a dominant criterion in determining how frequently dischargers should be inspected.

However, to the extent that past behavior of dischargers is correlated with future behavior, basing inspection frequencies on past behavior is another potentially useful means of targetting scarce enforcement resources. Thus, it may be fruitful to inspect more frequently those dischargers that have a history of noncompliance, and give less attention to dischargers that have proved to consistently satisfy their permit requirements. The linkage between inspection frequency and past behavior could take a variety of forms and it could be specified by an appropriately constructed formula or be based on a less formal and more subjective scheme. Regardless of the method used for linking inspection frequency to past noncompliance, it would still be necessary to at least occasionally inspect all dischargers regardless of their compliance records, in order to provide them with an incentive to remain compliant.

More Vigorous Enforcement of Self-Monitoring/Reporting Requirements

Even if monitoring resources are better targetted, the sheer number of dischargers and the constraints on state and federal enforcement resources imply that self-monitoring and reporting will continue to be the backbone of the compliance monitoring program. At present, it appears that far more attention has been given to taking enforcement action against effluent limit violations than to self-monitoring/reporting violations. For instance, relatively detailed guidelines have been developed for assessing penalties for effluent limit violations, but analogous guidelines have not been developed for self-monitoring/reporting violations. Although it is true that effluent limit violations are the ultimate objects of concern, self-monitoring/reporting violations are no less important since they are likely to conceal effluent limit violations (a discharger faced with even a minimal penalty for failing to submit a discharge monitoring report would presumably submit a report if it were compliant with all effluent limits).

Given the position of self-monitoring/reporting in the overall enforcement program it is important that this deficiency be remedied. Detailed and easy-to-use guidelines should be established for penalizing monitoring/reporting violations. This alone would foster more vigorous enforcement of monitoring/reporting requirements by making it easier for regional and state authorities to assess penalties for monitoring/reporting violations. In addition, the Agency as a whole should make a commitment to more actively pursue penalties for failing to submit monitoring reports.

If firms are to have the proper financial incentive to report violations, an appropriate relationship must be maintained between the penalty for not reporting effluent limit violations and the penalty for the effluent limit violation itself. In general, the penalty for not reporting must be several times higher than the penalty for the effluent limit violation. This relationship should be considered when developing guidelines for monitoring/reporting violation penalties.

As in the case of inspection frequencies, present and future frequencies of self-monitoring/reporting could be routinely linked to the accuracy, completeness, and punctuality of past reports, as well as the extent and frequency of actual effluent limit violations. Thus, a discharger that has consistently satisfied monitoring/reporting requirements and has not substantially exceeded effluent limits would be required to submit self-monitoring reports less frequently than dischargers that have not done so. Not only does this reward past compliance and cooperation by the discharger, but it also serves to conserve resources on the part of both the discharger and the enforcement authorities who are required to process and analyze the reports. It should be noted, however, that this would require major changes in the NPDES permit.

This suggestion is once again based on the premise that past behavior of dischargers is likely to be a good indicator of future behavior. Any scheme for determining reporting frequency would have to be flexible enough to accomodate cases where evidence suggests that this premise may not be valid.

For example, a discharger with a history of compliance that has within the recent past altered production processes or undergone a change in management, would be required to increase its frequency of self-monitoring/ reporting.

Refining Existing Penalty Policy for Effluent Limit Violations

The difficulty of accurately quantifying the extent of effluent limit violations inevitably complicates the penalty determination process. However, even in cases where violations are easily measured, our case studies suggest that current penalty policy as actually practiced may not provide firms with a clear financial incentive to comply with effluent limits.

Existing EPA penalty policy states that penalties should recoup the benefits to the discharger from noncompliance and, in addition, should include an amount reflecting the gravity of the violation. In practice, however, given the difficulty of placing a dollar value on the damages resulting from violations (the gravity component), the focus of penalty determinations is on the benefits to the discharger of noncompliance. Penalty assessments commonly do not exceed the (full) benefits to the firm of noncompliance. However, the model indicates that a penalty equal to the benefit of noncompliance is unlikely to provide dischargers with the necessary financial incentive to comply with effluent limit requirements. More precisely, the analysis indicates that if all effluent limit violations are not detected and penalized with certainty, a penalty set equal to the benefits enjoyed by the firm from noncompliance will not deter violations. This result can be made intuitive by considering the following highly simplified example.

Suppose that a discharger's monthly cost of complying with its NPDES effluent limits is \$1,500 and the penalty the discharger would face if it did not incur any of these costs is \$1,502, which is the benefit from noncompliance (i.e., the compliance costs avoided) plus a minimal gravity component. In deciding whether or not to incur the \$1,500 and comply with its effluent limits, the discharger would take into account the likelihood of being caught and fined if it does not incur the compliance costs. If the discharger perceives that it will be caught and fined each and every time it fails to incur the necessary compliance costs, it is in the discharger's financial interest to comply with the effluent limits since the penalty exceeds the compliance cost.

In practice, however, it is unlikely that each and every violation will be detected. Even if current violations are detected, past violations are difficult to verify. Suppose that only one of every two violations would actually be detected and penalized. On average, the firm would expect to pay a penalty of \$751 for each month it does not incur the costs associated with meeting its effluent limits (0.5 times \$1502). This is smaller than the \$1500 cost of compliance. It would therefore be in the discharger's financial interest not to incur the compliance costs and simply pay the penalty whenever it is caught violating its effluent limits.

This example illustrates that the penalty must be adjusted for the likelihood or probability that a discharger will be caught and fined for violations if it is to have the desired deterrent effect. Under existing policy, however, there are no provisions for doing so. Although it is difficult to determine the probability with which firms are caught and fined for violations, further attention must be given to developing methods for estimating this likelihood so that it can be more easily incorporated in penalty determinations.

4.1 REVIEW OF THE LITERATURE

The first formal economic analysis of noncompliance and enforcement was presented by Becker (1968). Using a fairly simple model, Becker examined several hypotheses regarding criminal behavior and the socially optimal deterrence of crime (broadly defined). Although Becker's model and analysis have been shown to be flawed¹, his paper generated considerable interest among economists and focused attention on the economic aspects of enforcing rules and regulations.

A detailed description of Becker's model is not warranted here given its abstractness and limited relevance to the analysis of pollution control enforcement. However, a brief sketch of its relevant features would be appropriate, since the model is the point of departure of much of the subsequent economic literature on enforcement. Moreover, the model of optimal enforcement presented in Section 4.2 is very similar to Becker's in terms of its general structure.

As noted above, Becker's model is highly stylized. The only two policy variables contained in the model are: (1) the probability of paying a fine for committing an offense or crime, and (2) the magnitude of the fine. Implicit in this formulation is the assumption that all penalties have some monetary (i.e., fine) equivalent. For example, it is assumed that a prison term has a fine equivalent; thus, a person would be indifferent between, say, a month in jail and a fine of \$3,000.

The probability of an offender paying a fine is a composite of the probabilities of detecting the offense, catching the perpetrator, assessing the fine, and collecting it. By varying the probability of paying a fine and the magnitude of the fine, the government (or other relevant body) can control the number and magnitude of offenses -- the higher the fine or the probability of paying it, the lower the general level of offenses.

¹ For an excellent critique of Becker's work see the article by Stern in the volume edited by Heinecke (1977).

The relevant enforcement parameter to the offender is assumed to be the expected fine, which is simply the product of the fine (f) and the probability of paying it (P).² Thus, an offender is assumed to be indifferent between a high fine with a low probability and a low fine with a high probability, as long as the magnitude of the expected fine (pf) is the same. So, for example, probability/fine combinations of $(0.1/\$100)$ and $(0.5/\$20)$ would be perceived as equivalent by an offender, because both yield an expected fine of $\$10$ ($0.5 \times \$20 = 0.1 \times \$100 = \$10$).

The government's assumed objective when setting the probability and fine for an offense is to minimize the sum of the social costs of deterring the offense and the social damages associated with the offense. The two major policy-oriented conclusions of Becker's analysis of this model are that:

- It is generally not socially desirable to completely eliminate crime; and
- The economically optimal fine is equal to the offender's wealth.

Both these conclusions are rather striking and merit some explanation. In the case of the first conclusion, Becker introduces the notion of an "optimal level of crime", at which the marginal, or incremental, costs of deterring crime are balanced by the marginal social damages associated with crime. In general, the optimal level is one at which some offenses are tolerated because the costs of totally eliminating them exceed the damages they generate.

The economic rationale underlying Becker's second conclusion is somewhat more involved. The conclusion follows from Becker's argument that, from a social standpoint, raising the probability of paying a fine is socially costly because it entails devoting more resources to apprehending offenders, gathering evidence, and so forth whereas raising a fine is virtually costless since it merely represents an increased transfer payment from the offender to society, via the government. Although Becker concedes that there are real resource costs associated with collecting fines, he contends that these are largely independent of the magnitude of the fine: the cost of collecting a large fine is not much greater than the cost of collecting a small fine. Given this argument, it is clear that the least cost way to achieve any desired expected fine level is to set the fine as high as possible and then adjust the probability of paying the fine until the desired expected fine is achieved. The upper limit on the fine is, of course, the individual's wealth, since this is the maximum the individual could pay (assuming that all penalties are monetary). This, of course, also sets the upper limit on the expected penalty, since the probability of detection cannot be greater than one.

² Formally, the assumption is that offenders are risk-neutral. Although Becker's analysis is by no means predicated on this assumption, it is used here to simplify exposition of his model and results.

Although Becker's first conclusion is widely accepted, his second conclusion has been the target of considerable criticism, both on ethical and economic grounds. Many have argued, for example, that setting the fine for, say, double parking equal to an offender's wealth, as Becker's analysis suggests, is ethically unacceptable. Indeed, Becker's analysis suggests that the fine for virtually any offense, regardless of its nature and the social damages it causes, should equal the offender's wealth. However, as Stigler (1970) has pointed out, fines should be set so as to preserve what he terms "marginal deterrence". That is, fines should be set so as to provide greater deterrence for a serious offense than for a less serious offense. Stigler argues that setting fines for all offenses equal to an individual's wealth would be inconsistent with preserving marginal deterrence.³ For example, if the fine for failing to install pollution control equipment is identical to the fine for improperly operating installed equipment, a firm would choose not to install equipment.

Other analysts have criticized Becker's second conclusion on the grounds that there are significant costs to collecting higher fines, contrary to what Becker assumes. It has been argued, for instance, that higher fines are likely to induce offenders to engage more heavily in avoidance activities, such as tying up the relevant government agency in legal maneuvers or bringing political pressure to bear on the agency. As McKeen (1980) has pointed out, outlays by industry on such activities over the past decade have been considerable, and have undoubtedly raised the resource costs to society of enforcing environmental regulations.

Given the above criticisms of Becker's (crucial) assumption regarding the costliness of raising fines, it is apparent that the optimal fine will, in general, not equal the offender's wealth. Instead, its value will be determined by the relative costs of raising the probability of the fine and raising the fine itself. This issue is explored further in Section 4.2.3.

Since the path-breaking work of Becker and Stigler, numerous articles and books have appeared on the subject of noncompliance and **enforcement**.⁴ Unfortunately, much of this work is of limited relevance to the problem of enforcing pollution control regulations. However, a small body of literature has accumulated that specifically addresses the problem of noncompliance and enforcement in the context of pollution control regulations. The earliest work in this category is by Downing and Watson (1974). Using a detailed simulation model, they examine the effects of alternative pollution control

³ Although Stigler's argument seems quite reasonable on the surface, it is not entirely correct since deterrence is determined by the expected fine and not just the fine. Since the expected fine is equal to the fine times the probability of paying it, marginal deterrence could be preserved even if the fines for offenses are identical by varying the probability of paying the fine for different offenses.

⁴ For a recent survey of much of this literature, see Pyle (1983).

and enforcement policies on particulate matter emissions from coal-fired power plants. Although their results are specific to the problem they examine, and not particularly relevant to the issues addressed here, their study is noteworthy for its empirical focus. Indeed very little quantitative, empirical work on noncompliance and enforcement has been done since their study.

The first theoretical paper on noncompliance and enforcement was published by Harford (1978). The paper presents a model of a noncompliant, risk-neutral firm under two different pollution control policies: effluent limits and effluent taxes. A similar model is developed by Storey and McCabe (1980) for the case of a risk averse firm. Both sets of authors demonstrate that firms are more compliant when the probability of detecting and punishing violations and/or the penalty for violations is raised. (This result is derived and explained in Section 4.2.) However, neither Harford nor Storey and McCabe address the problem of determining the socially optimal probability and penalty.

More recently, Beavis and Walker (1983) developed a model of a market for transferable pollution rights in which firms are noncompliant. Their model is notable in that it explicitly addresses the fact that discharges are (frequently) stochastic (i.e., the quantity and composition of discharges are partly determined by random events, such as equipment malfunctions, that are beyond the control of the discharger). However, because the focus of their paper is on markets for transferable pollution rights, their results are of limited relevance to existing pollution control policies.

In addition to the above papers, specific mention should be made of the research done at Resources for the Future (RFF) on pollutant discharge monitoring (Russell, Harrington, and Vaughan 1985). The emphasis of this research is on the use of statistical quality control techniques for monitoring pollution discharges that are stochastic in nature. The research does not explicitly address the problem of optimal enforcement; in particular, it gives limited attention to the structure of optimal penalties. Although the results obtained by the RFF team are not promising, the research is unique in attempting to provide operational solutions to the problem of designing effective monitoring strategies.

4.2 A SIMPLE MODEL OF OPTIMAL ENFORCEMENT

This section presents a simple model of optimal enforcement policy that attempts to capture the salient features of the various models of enforcement presented in the economic literature. The model differs from those presented in the economic literature in that it is specifically tailored to the problem of enforcing environmental regulations, in particular, regulations that limit pollutant discharges.

Before presenting the model of optimal enforcement, a model of a noncompliant firm is developed and used to establish the relationship between the enforcement policy and the firm's level of noncompliance. This

relationship is then used in developing the optimal enforcement model. It is important to remember, however, that the focus of this chapter is on firms that require financial incentives to comply with environmental regulations: Thus, firms that elect to comply even in the absence of penalties or other inducements are not described by the model developed in this section. Only firms that strictly maximize profits are the focus of this model.

4.2.1 Model of a Noncompliant Firm

Enforcement of regulations is a complex process not only because the sets of instruments and methods available to authorities are numerous and interrelated, but also because conducting enforcement policy occurs within the context of firm profit maximization. That is, for firms that require financial incentives to comply with regulations, enforcement policy must take into account the fact that these firms make choices and alter their behavior in response to changes in enforcement policy itself. Hence, the only way to understand how changes in enforcement policy (changing penalties or altering the perceived probability of detection and penalization) can better achieve enforcement goals is to first characterize how this subset of firms respond to different financial incentives. Hence, this section presents a stylized model of how these types of firms decide the degree to which they will comply with environmental regulations, based only on the financial incentives offered by enforcement policy.

The extent to which firms violate a regulation, if they do so at all, depends on the relative magnitude of the compliance costs and the expected penalties for noncompliance that they **face**.⁵ Thus, the subset of firms that will not comply with regulations in the absence of financial incentives to do so tend to weigh the expected costs associated with being detected and fined for their noncompliance against the costs of compliance. The expected penalty for noncompliance is simply the product of the perceived probability of being caught and fined for a violation and the **fine**.⁶ Although often overlooked in the policy literature, the perceived probability of catching and fining violators is as important as the magnitude of the fine. For example, if no resources are devoted to detecting noncompliance and firms know this, a fine, no matter how large, will have no deterrent effect since the perceived probability of being fined will probably be zero.

This implies, as a consequence, that enforcement policy is composed of two integral parts: (1) the fines and penalties for noncompliance (penalty policy), and (2) the level of enforcement activities (enforcement strategy). Since the subset of firms that require financial incentives to induce them to comply with environmental regulations respond to the expected penalties, their

⁵ We assume, for simplicity, that firms are risk-neutral.

⁶ Throughout this section, the terms "penalty" and "fine" are used interchangeably. This implicitly assumes that all penalties have monetary equivalents.

perceptions of the probability of being caught and penalized are just as important as the penalties themselves. Deterrence depends centrally on firms' perceptions of what penalties and probabilities of detection and penalization are, not necessarily the actual objective probability. Hence, to the extent that the probability of enforcing against a noncompliant firm is believed to be more probable than it actually is, the expected penalties anticipated by firms are higher than one might otherwise think. Together the penalties and the perceived probabilities form the expected penalty to which firms respond in their compliance decisions. Both penalty policy and enforcement strategy are therefore included in the expected penalty.

Enforcement Policy and Noncompliance

In general, noncompliance with regulations can range from small violations to very large ones. For instance, a firm facing a regulation limiting total discharges of mercury to 10 grams per day might comply with the regulation and never discharge more than 10 grams per day, exceed the limit by a relatively small amount and discharge, say, 11 grams per day, or exceed it by a wide margin and discharge 100 grams a day. We shall refer to the difference between the actual amount the firm discharges and the amount it is allowed to discharge as the violation size. In the example just presented, the three potential violation sizes are: 0, 1, and 90 grams.'

Firms that require financial incentives to comply with the regulation will tend to comply with regulations only to the point at which it minimizes the sum of the compliance costs and expected penalties faced. Exhibit 4-1(a) depicts a firm's compliance costs and expected penalties as a function of violation size. The monetary value of the damages avoided are also depicted (although, by assumption, these do not influence the firm's compliance level since we are dealing here with firms that do not wish to comply with regulations except to the extent that financial incentives exist for them to do so).

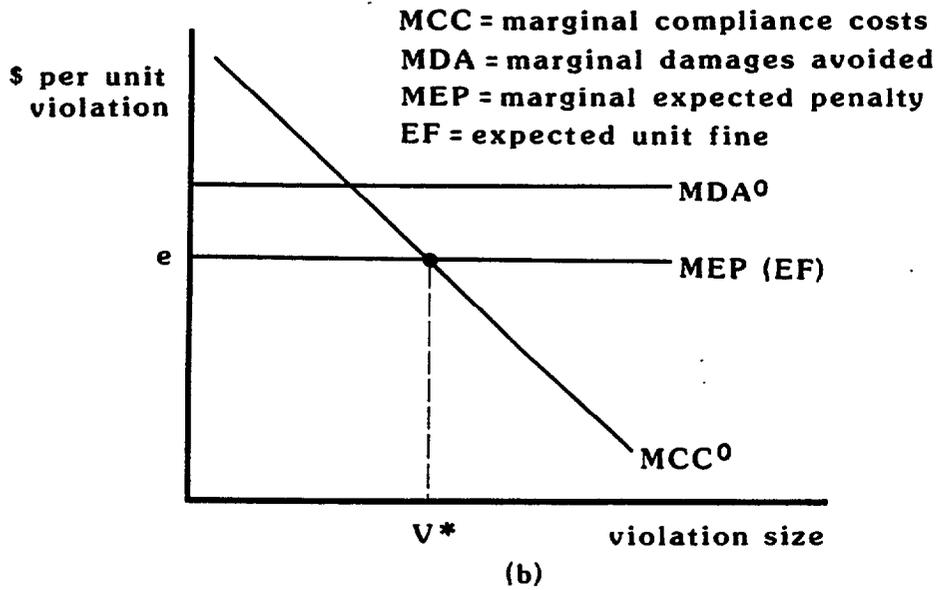
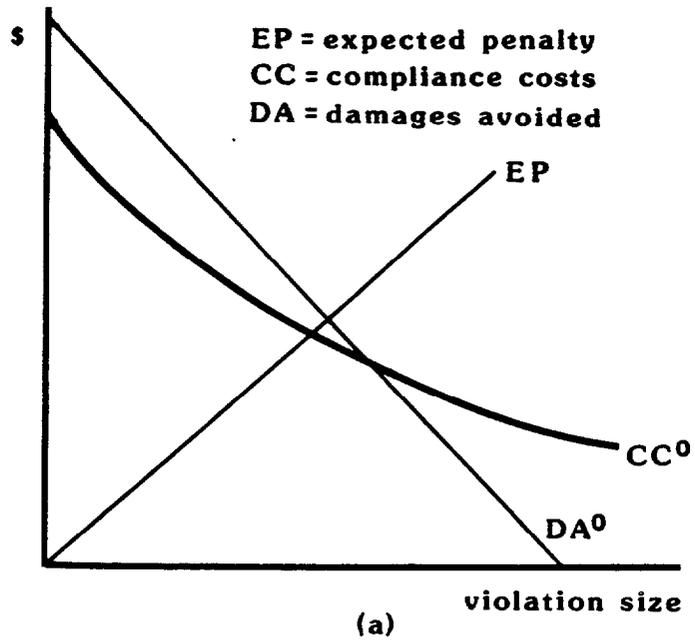
As shown in the figure, the firm's compliance costs (CC^0) diminish as the violation gets larger (since less control costs less), whereas the expected penalty (EP) rises as the violation gets larger. Larger violations are assumed to pose larger penalties, so in the face of a constant perceived probability of detection and penalization, the expected penalties firms perceive rise with larger violations.

The dollar value of the damages avoided (DA^0) also diminish as the violation gets larger. Damages avoided are calculated by subtracting the damage associated with a particular violation size from the damage associated with the maximum violation size that a firm would choose. For example, if the.

⁷ In this simple model, we abstract from the possibility that violations may be determined by averaging several days' effluent information, as well as other complexities of the enforcement-compliance process.

Exhibit 4-1

Noncompliant Firm's Violation Size



damage associated with the maximum violation size is \$100, and the damage associated with a violation size of, say, 2 grams, is \$90, then the damage avoided given a violation of 2 grams is \$10 (\$100 - \$90). Assuming that damages increase with the size of the violation, damages avoided decrease as the violation gets larger.

Exhibit 4-1(b) shows the "marginal curves" corresponding to the "total curves" in Exhibit 4-1(a). The marginal compliance cost curve (MCC^0) slopes downward, which is consistent with both intuition and empirical observation: the unit cost of abating pollution rises as less and less pollution is generated (i.e., as the violation size gets smaller and **smaller**).⁸ For example, reducing a violation by one gram is generally cheaper when the initial violation is 90 grams than when it is two grams. The downward slope of the MCC^0 curve also suggests one of its central roles in the analysis (i.e., it measures the marginal benefit to the firm of lower compliance, since these are expenditures avoided).

The marginal expected penalty curve in Exhibit 4-1(b) (MEP) is flat given the simplifying assumption that total expected penalties increase linearly with violation size. As long as the perceived probability of detection and penalization is constant, under this linearity assumption, each unit of additional violation raises the expected penalty by the perceived probability of detection and penalization times the marginal penalty per unit of violation. The expected fine per unit violation is therefore constant (e.g., \$100 per gram of mercury over the allowed daily level).

The marginal damages avoided curve (MDA^0) is also flat given the simplifying assumption that damages avoided decrease linearly with violation **size**.⁹ This implies that the monetary value of the damage resulting from each unit of pollution is constant, and does not vary with the total amount of pollution generated.

The firm's degree of noncompliance is given by v^* , which is the violation size at which the marginal saving in compliance costs is equal to the marginal expected penalty. The reason why a violation of this size minimizes the sum of the firm's compliance costs and its expected penalties is clear from Exhibit 4-1(b): as a violation larger than v^* , the incremental savings in compliance costs are smaller than the incremental costs in the form of what the firm believes to be the higher expected penalties it faces. Conversely, at a violation smaller than v^* , the incremental savings, in terms of what the

⁸ The curve labeled MCC^0 actually gives the negative of marginal compliance costs and should therefore be labeled $-MCC^0$; however, in order to maintain consistency with later figures, and to avoid confusion, the minus sign is omitted in the figure. This caveat also applies to the marginal damages avoided curve.

⁹ This assumption does not affect our conclusions.

firm believes to be the lower expected penalties, are smaller than the incremental compliance costs. From this it should be clear why it is the firm's perception of penalties and the probability of detection and penalization that matters for determining its degree of compliance, not the actual underlying objective probability. It is perceptions that motivate firms' choices, hence the perceived expected penalty matters for enforcement policy.

The model of the noncompliant firm assumes implicitly that firms are not routinely forewarned of on-site inspections. Thus, firms are assumed to face at least some uncertainty about when they are inspected. In practice, however, firms are frequently notified in advance of upcoming inspections. This affects our analysis only if firms are able to easily reduce their discharges on short notice by shutting down production processes and stepping up treatment of effluents. If this is feasible, firms could quickly bring themselves into compliance during on-site inspections, making it difficult for the agency to detect violations.

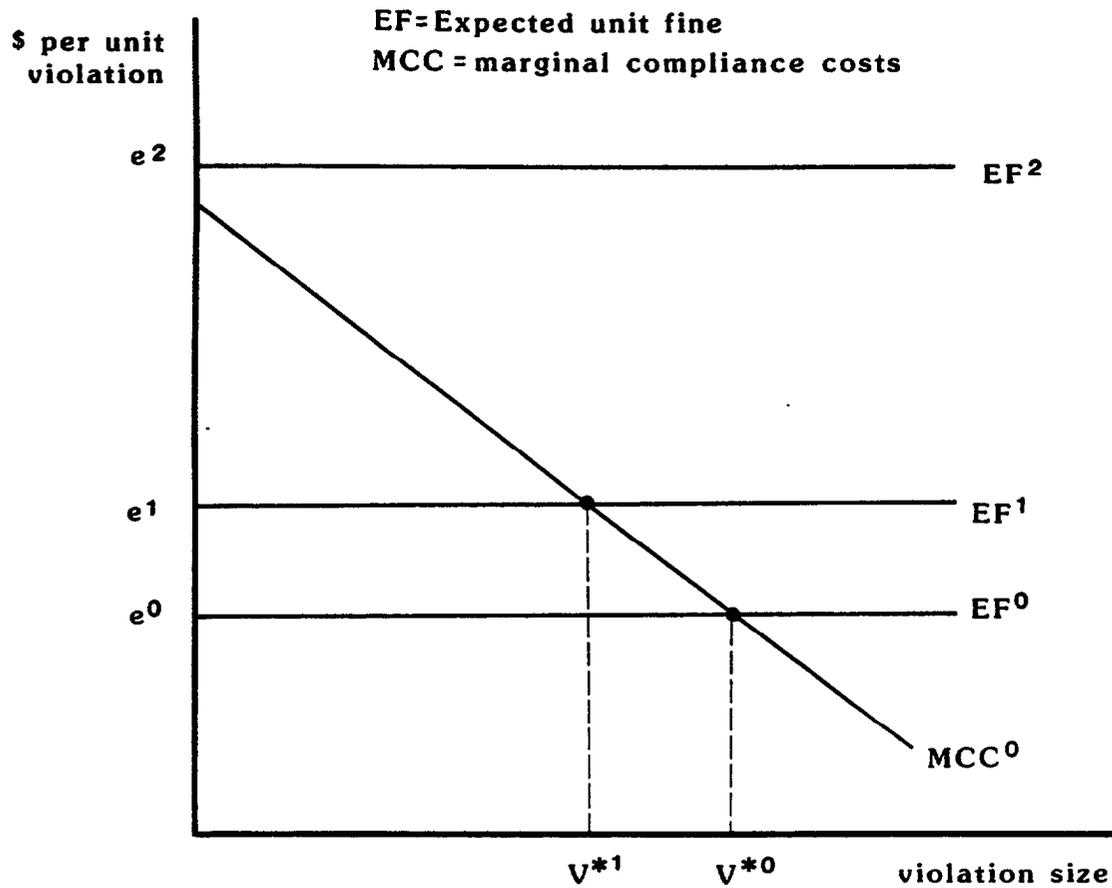
In general, the feasibility of this sort of strategy depends on the production and treatment technologies involved and the amount of advance notice the firm is given. Standard biological treatment processes for wastewater have start-up times of several days, which makes it difficult to step up treatment on short notice. Moreover, even if it is technically feasible to reduce pollution discharges on short notice, the associated costs may be high enough to dissuade firms from doing so.

Effects of Increases in the Expected Fine and on Marginal Compliance Costs on the Degree of Noncompliance

The relationship between a firm's violation size and the expected fine it faces can be determined by examining the effect of an increase in the expected fine on the firm's violation size. The effect of such an increase is illustrated in Exhibit 4-2. As shown, an increase in the expected fine from e^0 to e^1 lowers the firm's violation size from v^{*0} to v^{*1} . Since the expected fine (pf) is equal to the fine per unit violation (f) times the perceived probability of being caught and fined (p), this result implies that an increase in either the perceived probability (p) or the fine (f) would lower the firm's violation size (provided the firm requires financial incentives to comply with regulations) because an increase in either variable would raise the expected fine (pf). Thus, enforcement policy can reduce violation sizes by either increasing penalties or increasing the perceived probability of detection and penalization, since these together form the expected penalty.

If the expected fine is large enough, the firm will be perfectly compliant and set its violation equal to zero. This is also shown in Exhibit 4-2. When the expected fine is set at e^2 , it does not pay the firm to be noncompliant, because, for even the smallest violation, the savings in terms of lower compliance costs are smaller than the costs in terms of higher expected penalties. This would also be true for any expected fine higher than e^2 .

Exhibit 4-2
Effects of Increases in the Expected Fine on the Firm's Violation Size



Again, the expected fine can be raised by increasing either penalties or firms' perceptions of the probability of detection and penalization, as discussed in detail below.

The relationship between a firm's marginal costs of compliance and its violation size can be determined similarly by examining the effect of an increase in compliance costs on the firm's compliance rate. Exhibit 4-3 illustrates the effect of an upward shift in the marginal cost of compliance (MCC^0) curve on the firm's violation size. As the figure shows, an increase in marginal compliance costs from MCC^0 to $MCC^{0'}$ raises the firm's violation size from v^{*0} to v^{*1} .

Thus, the two main conclusions of the analysis of the noncompliant firm are:

- The violation size of a noncompliant firm falls when the expected fine it faces goes up; and
- The violation size rises when the firm's marginal costs of compliance increase.

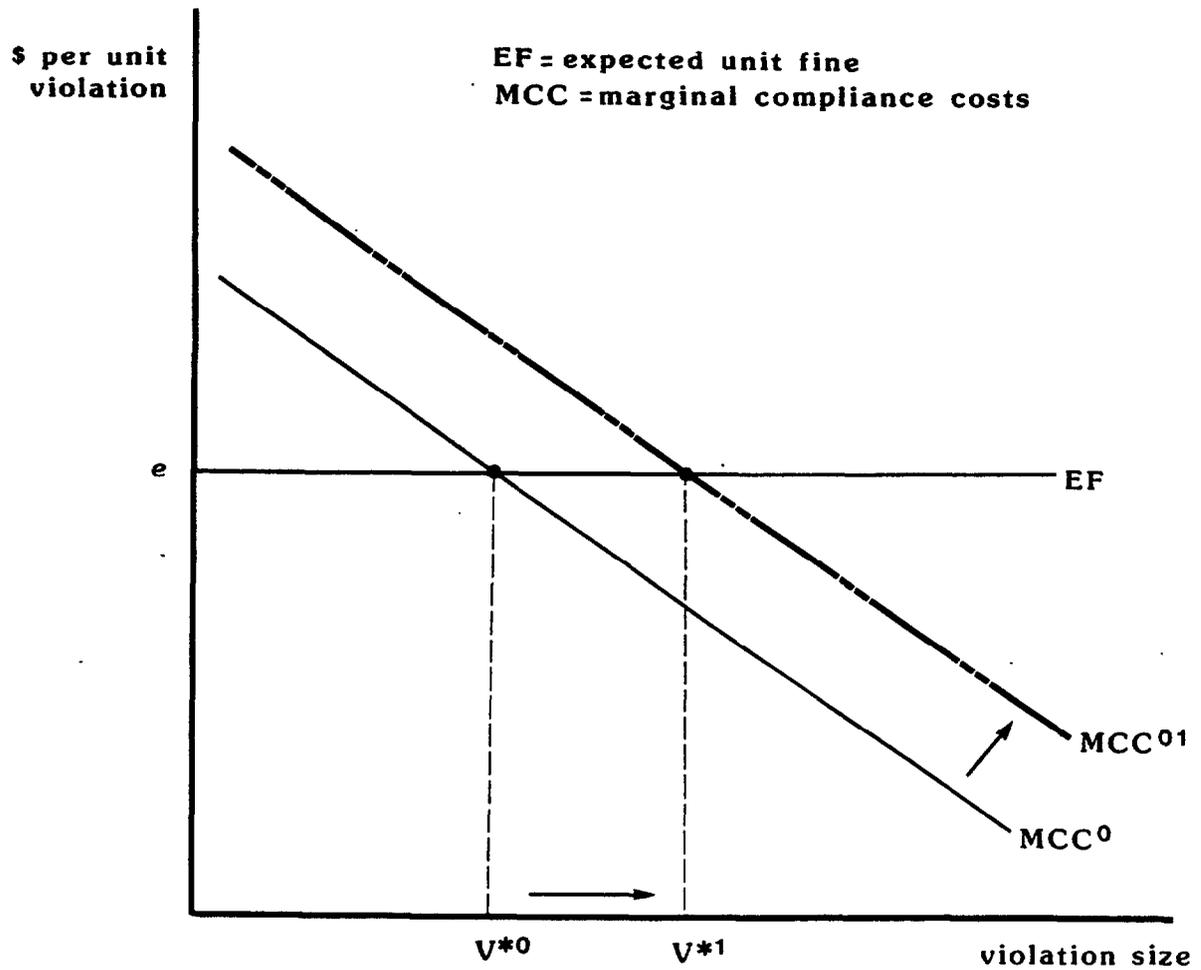
This section has outlined how the subset of firms that require financial incentives to comply with regulations respond to changes in both enforcement policy (i.e., changes in penalties and changes in perceived probabilities of detection and penalization) and to changes in compliance costs. We now turn to investigate how the enforcement authorities should respond to firms' choices to optimally enforce environmental regulations given limited enforcement resources.

4.2.2 Model of Optimal Enforcement

Determining the optimal level of enforcement requires an evaluation of the social benefits and costs associated with enforcement. The obvious direct costs of enforcement are the resources expended on monitoring firms' discharges, gathering evidence on violations, and assessing penalties for noncompliance. But in addition to these direct costs, there are indirect costs associated with enforcement. As discussed above, increased enforcement (i.e., a higher expected fine) induces firms to increase compliance. This implies that firms spend more on pollution control. These increased compliance costs must also be considered in a comprehensive benefit-cost analysis of enforcement even though they are borne by firms that are noncompliant. Thus, there are two types of costs associated with enforcement: the cost of enforcement (enforcement costs) and the expenditures by firms on compliance due to increased enforcement (compliance costs). The total social cost associated with enforcement is given by the sum of enforcement costs and compliance costs.

The social benefits of enforcement result from the increased compliance of firms when the level of enforcement is increased. Increased compliance implies lower damages in terms of human health and environmental quality due to pollution. The central problem of optimal enforcement is trading off these social benefits from increased enforcement against the attendant costs.

Exhibit 4-3
Effect of Higher Marginal Compliance Costs on Firm's Violation Size



It is important to note that in the context of the model, the level of enforcement is directly related to the magnitude of the expected fine: the greater the resources devoted to enforcement, the higher the expected fine and, hence, the greater the incentive for compliance. Thus, if one exacts higher penalties, the expected fine anticipated by firms increases. Similarly, if one devotes a greater amount of enforcement resources to increasing the perceived probability that firms will be detected and penalized, this too will raise the expected fine.

Hence, the problem of determining the optimal level of enforcement reduces to one of determining the optimal value of the expected fine. However, it is important to remember that the expected fine involves much more than simply setting penalties. Setting the perceived probability that detection and penalization will occur is no less important in the overall enforcement process. Hence, although it is correct to claim that optimal enforcement policy reduces to setting the optimal level of the expected fine, one should keep in mind that this is a much broader mandate than setting penalties alone. Indeed, one of the most difficult aspects of enforcement policy is trying to decide exactly how much should be spent, in what areas, and on what activities to ensure that the perceived probability of detection and penalization is sufficiently high.

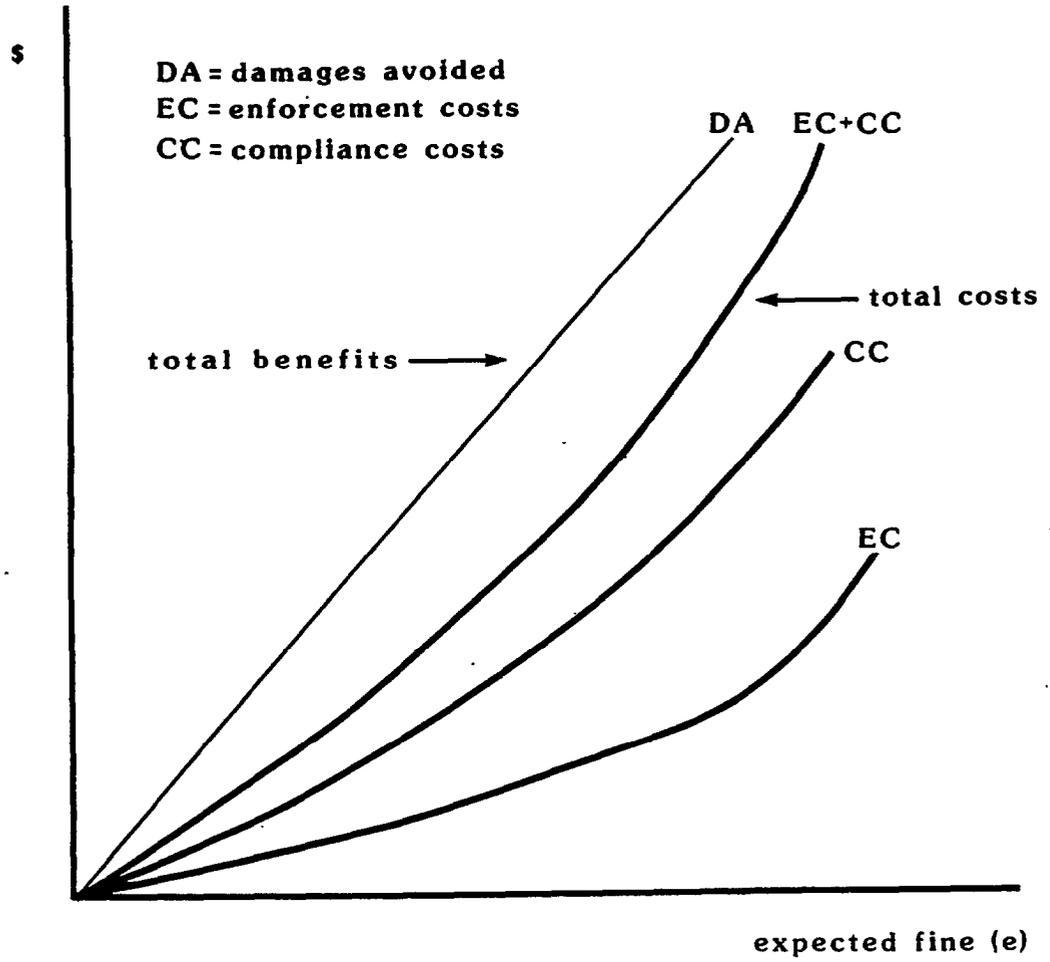
Exhibit 4-4 depicts the social costs and benefits associated with enforcement. The costs of enforcement itself are given by the curve labeled EC. As shown, enforcement costs rise as the expected fine, e , gets larger since more resources must be devoted to catching and fining violators. The curve EC is drawn, however, for a fixed relationship between the amount of resources spent on enforcement activities and the resulting perceived probability of detection and penalization. As shown below, if this relationship were to change over time, or if it differs depending on the industry investigated, this would imply that the position of the EC curve changes as well.

The compliance costs associated with enforcement are given by the curve labeled CC. The curve is upward sloping because as the expected fine perceived by firms increases, firms increase their compliance, which implies that they spend more and more on pollution control. The total social costs of raising the expected fine are given by the sum of enforcement costs (EC) and compliance costs (CC). These are represented in Exhibit 4-4 by the curve labeled EC+CC, which is simply the sum of the enforcement cost and compliance cost curves.

The total social benefits of raising the expected fine, on the other hand, are represented in Exhibit 4-4 by the curve labeled DA. As noted above, the benefits of enforcement are the damages to human health and environment that are avoided. The benefit curve is upward sloping because damages avoided rise as the expected fine goes up and firms increase their compliance.

Exhibit 4-4

Cost and Benefits of Enforcement



The Optimal Expected Fine

The socially optimal expected fine is the value of the expected fine at which net social benefits, the difference between total social benefits and total social costs, are maximized. This corresponds to the value of the expected fine at which the marginal benefits of increasing the expected fine equal the marginal costs of doing so. Exhibit 4-5 presents the marginal benefit and marginal cost curves corresponding to the "total curves" in Exhibit 4-4.

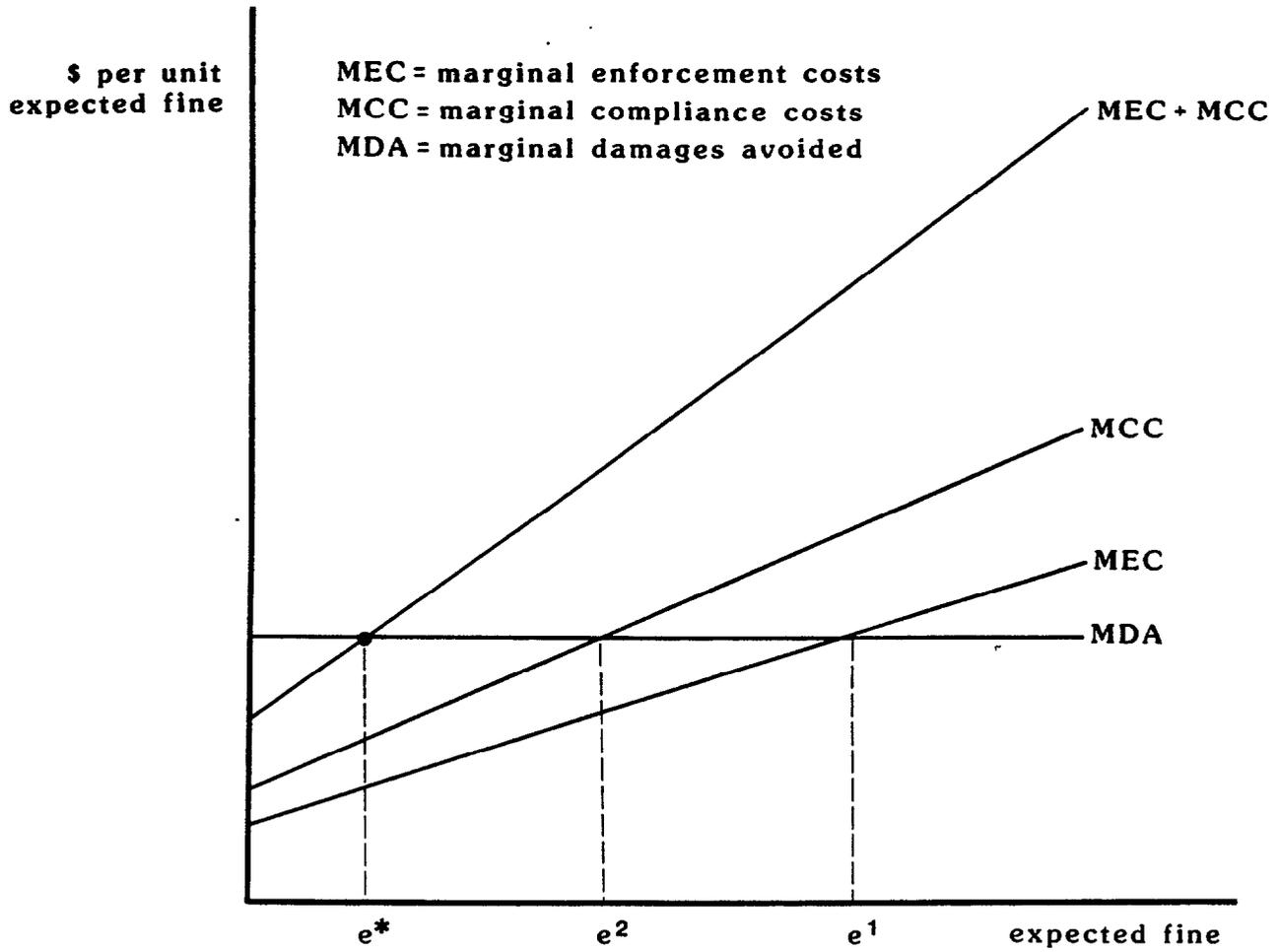
The marginal benefits are simply the marginal damages avoided, given by the curve labeled MDA. The marginal costs consist of two components: marginal enforcement costs (MEC) and marginal compliance costs (MCC); the sum of these two is given by the curve labeled MEC+MCC. The optimal value of the expected fine, e^* , is given by the intersection of MDA and MEC+MCC, which is the point at which marginal benefits equal marginal costs. As can be seen in Exhibit 4-5, if either the marginal compliance costs or the marginal enforcement costs are (incorrectly) ignored when choosing the optimal expected fine, it will be set too high and there will be too much enforcement. For example, if it is incorrectly assumed that the optimal level of the expected fine is given by the intersection of MDA and MEC, the expected fine will be set too high (e^1 is larger than e^*). Intuitively speaking, setting the expected fine such that the marginal damages avoided are equal to the enforcement costs only ignores the fact that real social resources must be devoted to compliance costs. Ignoring these additional costs incorrectly assumes that compliance is free from a social perspective. Similarly, if enforcement costs are ignored and the expected fine is set equal to the value at which MDA and MCC intersect (e^2), this would incorrectly assume that enforcement itself is costless from a social perspective, which is manifestly untrue.

The marginal compliance cost and marginal damages avoided curves in Exhibit 4-5 are unusual in that they are functions of the expected fine rather than of violation sizes or pollutant levels, as is typically the case in standard analyses of the benefits and costs of pollution control. Thus, these curves represent the changes in compliance costs and damages avoided resulting from an increase in the expected fine, unlike the more typical curves in Exhibit 4-1, which represent the changes in compliance costs and damages avoided due to a larger violation.

Moreover, a comparison of the marginal compliance cost curves in Exhibits 4-1(b) and 4-5 reveals that in one case (Exhibit 4-1(b)) the curve is downward sloping, while in the other case (Exhibit 4-5) it is upward sloping. These curves are actually consistent because two different marginal compliance cost curves are involved. The curve labeled MCC^0 in Exhibit 4-1(b) gives the marginal costs of compliance as a function of the firm's violation size, whereas the curve labeled MCC in Exhibit 4-5 gives the marginal costs of compliance as a function of the expected fine (compare the horizontal axes on the two figures). The MCC curve in Exhibit 4-5 is based on the MCC^0 curve in Exhibit 4-1(b). However it takes into account the endogenous relationship between the firm's degree of compliance and the expected fine the firm

Exhibit 4-5

The Optimal Expected Fine



faces.¹⁰ As explained earlier, higher expected fines induce smaller violations, which in turn imply higher marginal compliance costs (see Exhibit 4-1(b)). Marginal compliance costs therefore increase as the expected fine gets larger (see Exhibit 4-5):

higher expected fine \implies smaller violation \implies higher marginal
compliance costs.

The marginal damage curves in Exhibits 4-1(b) and 4-5 also must be interpreted with care. Typically, as in Exhibit 4-1(b), damages avoided are expressed as a function of violation size (or, equivalently, pollutant levels). However, in Exhibit 4-5, marginal damages avoided are presented as a function of the expected fine. The MDA curve in Exhibit 4-5 can be derived from the MDA^0 curve in Exhibit 4-1(b) by accounting for the endogenous relationship between the firm's optimal violation size and the expected fine it **faces.**¹¹ The MDA curve has the same slope as the MDA^0 curve only because the MDA^0 curve is flat. This implies that marginal damages are independent of violation size and, as a result, the relationship between optimal violation size and the expected fine has no influence on the shape of the MDA curve.

One final comment about the marginal enforcement costs (MEC) curve is necessary. MEC in Exhibit 4-5 is drawn for a given fixed relationship between enforcement activities and the resulting perceived probability of detection and penalization. If this relationship varied for different industries or areas of the nation (say a given level of perceived probability could be achieved with fewer expenditures of enforcement resources in certain locations), then the MEC curve would shift downward, suggesting that the optimal expected fine would be higher than otherwise. This implies that in these circumstances, the same level of enforcement activities combined with the same penalty structure would result in a larger expected fine and hence, a higher degree of compliance. More will be said below concerning the optimal composition of the optimal expected fine in terms of the size of penalties versus the size of the perceived probability of detection and penalization.

¹⁰ Formally, marginal compliance costs as a function of the expected fine ($MCC(e)$) are obtained by substituting the expression for the firm's optimal violation size as a function of the expected fine it faces ($v^*(e)$) into the expression for marginal compliance costs as a function of violation size ($MCC^0(v)$): $MCC^0(v^*(e)) = MCC(e)$.

¹¹ Formally, marginal damages avoided as a function of the expected fine ($MDA(e)$) are obtained by substituting the expression for the firm's optimal violation size as a function of the expected fine ($v^*(e)$) into the expression for marginal damages avoided as a function of violation size ($MDA^0(v)$): $MDA^0(v^*(e)) = MDA(e)$.

Determinants of the Optimal Expected Fine

It is clear from the above discussion that the optimal expected fine depends on four factors:

- marginal damages avoided (MDA);
- marginal enforcement costs (MEC);
- marginal compliance costs (MCC); and
- the relationship between firms' perceived probability of detection and penalization and enforcement activities.

However, the optimal expected fine does not depend on these four factors in any simple way. For instance, it is not equal to the sum (MDA + MEC + MCC), or (MDA - MEC - MCC). More importantly, the optimal expected fine does not depend in any simple way on marginal damages avoided and marginal compliance costs when these are expressed as functions of violation size, which is the usual way in which these factors are presented. As a result, computing the optimal expected fine is a difficult task that can only be accomplished with detailed knowledge of the firm's compliance costs, the damages from noncompliance, the costs of enforcement, and the relationship between the perceived probability and actual enforcement activities. Thus, the model presented does not provide a simple means of calculating the optimal expected fine.¹² However it does provide some general guidelines regarding the broad characteristics of optimal enforcement. These are developed below by examining the relationship between the optimal expected fine and each of the four factors listed above.

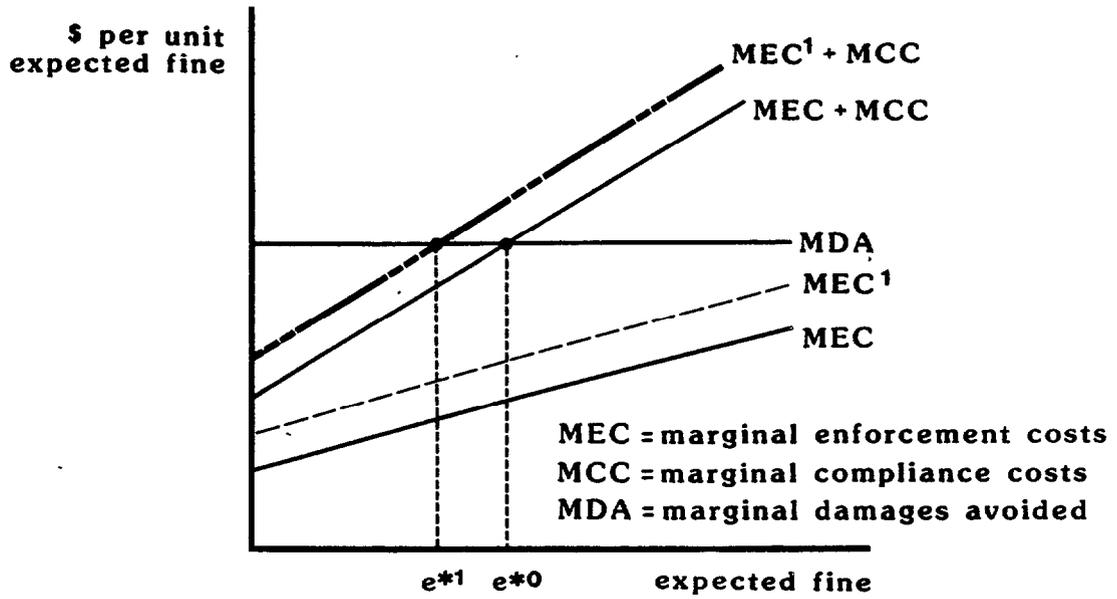
Effect of an Increase in Marginal Enforcement Costs on the Optimal Expected Fine

The relationship between the optimal expected fine and the magnitude of enforcement costs can be determined by evaluating the effect of an increase in marginal enforcement costs on the value of the optimal expected fine. This is illustrated in Exhibit 4-6. As shown, the upward shift in the marginal enforcement cost curve, MEC, also shifts up the total marginal cost curve, MEC+MCC, lowering the optimal value of the expected fine from e^*0 to e^*1 . This result is easily explained: an increase in marginal enforcement costs, holding everything else constant, implies higher total marginal costs of enforcement. Since marginal benefits are unchanged, fewer resources should be devoted to enforcement. Put simply, if the marginal benefits of enforcement are held constant, fewer resources should be devoted to enforcement the more costly it becomes.

¹² Given the complexity of the enforcement problem, no comprehensive model of optimal enforcement would yield an operational means of calculating the optimal expected fine.

Exhibit 4-6

Effect of Higher Marginal Enforcement Costs on the Optimal Expected Fine



Effect of an Increase in the Perceived Probability of Detection and Penalization on the Optimal Fine

Modeling the impact of an increase in the relationship between actual enforcement activities and firms' perceptions of the probability of detection and penalization can also be accomplished within this framework. In Exhibit 4-7, the curve MEC is drawn for a given relationship between perceptions and reality, so that the optimal expected penalty, given the marginal compliance cost and marginal damages avoided schedules, is e^{*0} (where MEC + MCC cross MDA). However, suppose that another set of firms (say, in a different industry) believe that the probability of detection and penalization given the same levels of enforcement activities is higher. In this event, the MEC curve relevant for these firms becomes MEC^1 , which lies below MEC since it takes less actual enforcement resources than before to achieve a given level of the expected penalty. This implies that the optimal expected penalty is higher for this second set of firms, or e^{*1} (the intersection of MEC^1 + MCC with MDA).

This result accords with intuition since, at least analytically, enhanced perceived probability has much the same impact as cheaper enforcement costs. In a very real sense, this enhanced perception of apprehension and penalization means that it is cheaper to achieve a given level of expected penalty. Furthermore, at least in the context of a constant level of marginal damages avoided, it is also apparent that the amount of enforcement resources actually devoted to enforcement falls as the relationship between perceptions and actuality is further exaggerated. That is, for otherwise identical situations or firms, if one has a far greater perception that detection and penalization will follow violations, then less enforcement resources are necessary to achieve an even higher level of the expected fine (and hence, compliance). This, of course, is consistent with a policy that sometimes uses "hit and run" enforcement tactics in certain geographical areas or for certain types of firms. In these cases, a relatively small expenditure can yield large results in terms of compliance, which makes this relationship of fundamental importance to deciding how best to target scarce enforcement resources. This may also provide a rationale for occasional enforcement actions in areas where typically no actions are taken.

Effect of an Increase in Marginal Damages Avoided on the Optimal Expected Fine

In studying the relationship between marginal damages avoided and the optimal expected fine, one has a choice of examining the effect of an increase in marginal damages avoided expressed as a function of either violation size (MDA) or expected fine (MDA). However, as argued earlier, shifts in the MDA^0 curve are mirrored by the MDA curve; therefore it does not matter which curve we shift. This is illustrated in Exhibit 4-8. As shown, an upward shift of the MDA^0 curve results in an upward shift of the MDA curve. The effect of this shift is to increase the optimal expected fine from e^{*0} to e^{*1} . Once again, the rationale underlying this result is intuitive: the upward shift in marginal damages avoided, holding everything else constant,

Exhibit 4-7

Effect of Increased. Perceived Probability of Detection and Penalization on the Optimal Expected Fine

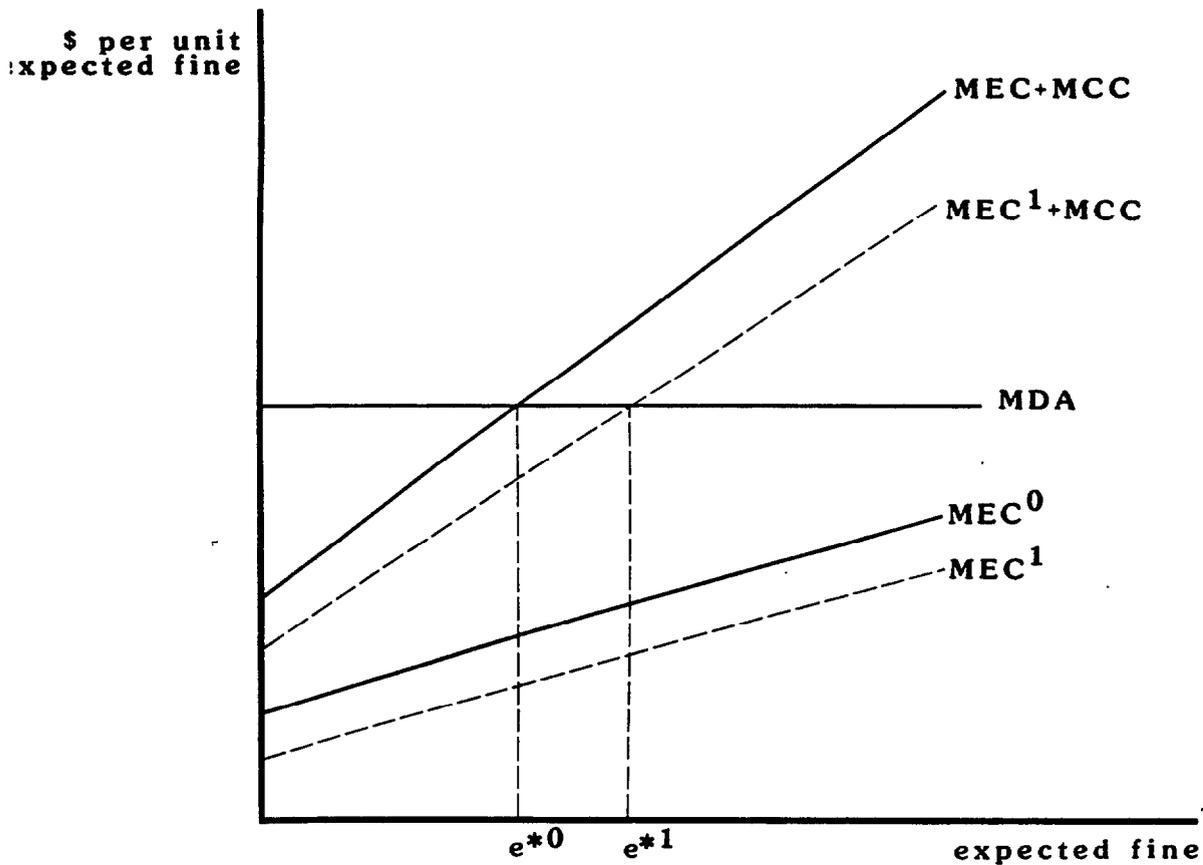
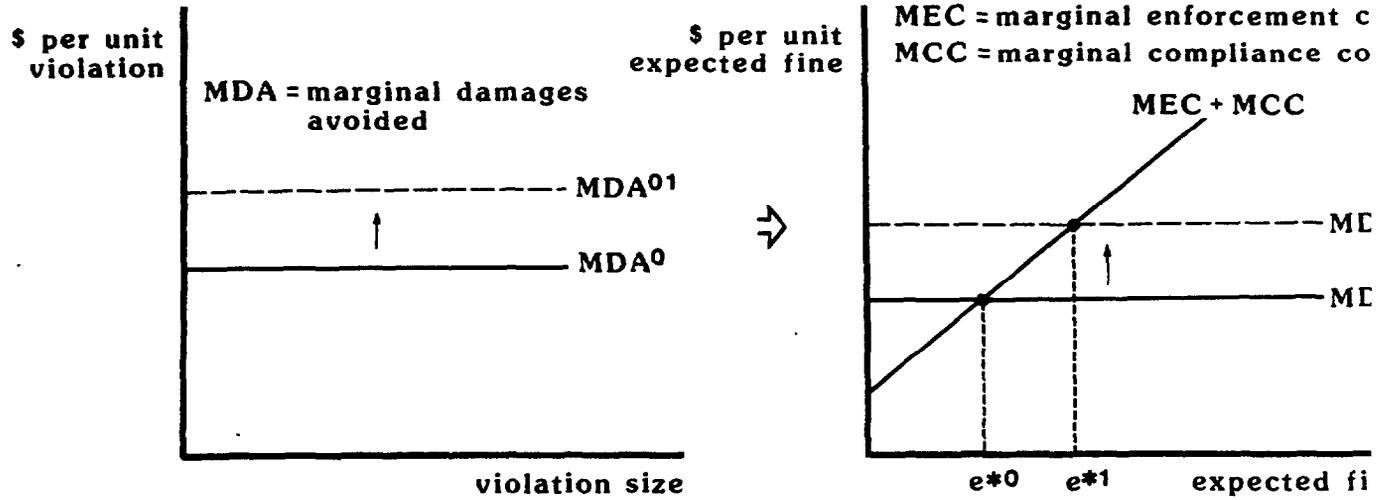


Exhibit 4-8

Effect of Higher Marginal Damages Avoided on the Optimal Expected Fine



raises the marginal benefits of enforcement without changing the marginal costs. As a result, more resources should be devoted to enforcement in this area relative to other situations.

Effects of an Increase in Marginal Compliance Costs
on the Optimal Expected Fine

The relationship between the level of compliance costs and the optimal settings of the fine and of enforcement efforts is both complex and . ambiguous. That is, higher marginal compliance costs could either raise or lower the level of the expected fine depending on the settings of other aspects of the optimization framework. Hence, this analysis is presented in an appendix to this chapter.

The conclusions we can draw from the foregoing analysis of the relationship between the optimal expected fine and the various determinants of its level are interesting and, for the most part, fairly intuitive. There are several contexts in which these conclusions can be understood. One, of course, is that these theoretical conclusions give policy makers an indication of what matters in enforcement policy. That is, the theoretical model and the conclusions one can draw indicate which elements are important to understand and, to the extent feasible, measure when developing both penalty policy and enforcement strategy. The other sense in which these conclusions can be of significant service is in helping to develop overall enforcement policies for different types of firms, sectors of the nation, and types of pollutants. Thus, the comparative statics of the model indicate how firms or situations that are manifestly different, in terms of the characteristics that affect the settings of the penalties and probabilities of detection and penalization, should indeed be treated differently in an optimized enforcement policy.

To review, the significant conclusions from the model of optimal enforcement are the following:

- The optimal expected fine depends, in a fairly complicated way on four factors: (1) marginal compliance costs, (2) marginal enforcement costs, (3) marginal damages avoided, and (4) the relationship between actual enforcement activities and firms' perceptions of the probability of detection and penalization, regardless of whether these are expressed as functions of violation sizes or expected fines. In particular, the optimal expected fine is not equal to the sum of these four factors.
- The optimal expected fine rises when marginal damages avoided increase (where marginal damages are expressed as a function of violation size or pollutant levels).
- The optimal expected fine falls when marginal enforcement costs rise.

- The optimal expected fine rises when perceptions of the probability of detection and penalization rise.
- The optimal expected fine may rise or fall when marginal compliance costs increase (where marginal costs are expressed as a function of violation size or pollutant levels).

The study develops the implications of these results for the design of enforcement policy further below. First, however, the composition of the optimum expected fine in terms of its two components -- the fine itself and its perceived probability -- is examined more closely.

The Optimal Values of the Fine and its Probability

The analysis presented so far has been couched entirely in terms of the expected fine. In this section we examine the relationship between the optimal expected fine and the optimal fine since much of the current debate on enforcement has focused on the optimal fine (penalty).

As made explicit earlier, the model of enforcement is based on the simplifying assumption that the marginal expected penalty or fine is constant. This implies that the expected fine per unit violation is constant. For example, the expected fine for exceeding a daily BOD limit is \$50/pound regardless of whether the violation is 10 pounds in excess of the allowed daily level or 100 pounds in excess. The expected fine per unit violation (e) is therefore simply equal to the perceived probability of catching and fining violators (p) times the fine itself (f): $e = pf$. The total expected fine is equal to the expected fine per unit violation times the violation: pfv . The total fine is simply the fine times the violation size: fv .

Our model of optimal enforcement provides, at least in principle, the optimal value of the expected fine (e^*), but it does not give the optimal values of the fine and its probability (as perceived from the perspectives of the firms in the regulated community); it only requires that the product of these two variables equal a specified value. For instance, it does not tell us whether a small fine should be applied with a high perceived probability, or a large fine with a small perceived probability.

To determine the optimal values of the the fine and its perceived probability, one must consider the relative costs of (increasing) the fine and the perceived probability. More specifically, one must determine the least cost combination of the fine and its perceived probability that gives an expected fine of e^* . Earlier, the review of the literature discussed Becker's argument that raising fines is socially costless, whereas raising the probability of catching and fining offenders is costly since it entails devoting more resources to monitoring firms and gathering evidence on violations. This argument led Becker to conclude that the least cost means of achieving any desired expected fine is to set the fine as high as possible (equal to the offender's wealth), and then adjust the probability until the desired expected fine is obtained.

However, as pointed out before, Becker's argument is erroneous. Higher fines induce offenders to devote more resources to avoiding being caught. In addition, higher fines imply stricter standards of evidence, thereby requiring more resources to be devoted to collecting evidence and developing a sufficiently strong case. Consequently, raising fines for violations is not socially costless.

Given that both higher perceived probabilities of catching and fining violators and higher fines are more costly to achieve, determining the optimal values of the fine and its probability requires detailed information on the relative costs of each. Exhibit 4-9 shows the relationship between relative costs and the optimal values of the fine (f) and its perceived probability (p) for a given value of the expected fine (e^*). The curve labeled EF gives the combination of values of p and f that yield an expected fine of e^* . At any point along this curve, the product of the corresponding values of p and f is equal to e^* ($pf = e^*$). The curve slopes downward because a smaller value of, say, p must be compensated for by a higher value of f if the expected fine is to remain at its initial value.

The family of curves labeled EC indicate the costs of setting the fine and its perceived probability at various levels, given a constant relationship between the perceived probability and the underlying actual enforcement expenditures. Thus, these curves represent the costs of enforcement. Along any given curve, such as E^1C^1 , enforcement costs are **constant**.¹³ This explains the downward slope of the curves: if enforcement costs are to remain constant, p must fall when f gets higher (or vice versa). As one moves to curves farther from the origin, enforcement costs increase because both the fine and its perceived probability increase. For instance, enforcement costs are higher along the curve labeled E^2C^2 than along E^1C^1 .

The least-cost combination of values of the fine and its perceived probability that yield an expected fine of e^* is determined by identifying the point at which the enforcement cost curve closest to the origin touches the expected fine curve EF. In Exhibit 4-9 this point is labeled Z, and the optimal (i.e., cost-minimizing) values of the fine and its perceived probability are f^* and p^* , respectively. The enforcement cost curve must touch the expected fine curve, otherwise no combination of values of p and f along the cost curve will yield an expected fine of e^* . Furthermore, we are interested in the least-cost means of achieving the expected fine, therefore, the cost curve closest to the origin is the relevant one.

The relative costs of raising the fine and raising the perceived probability determine the slope of the enforcement cost curves. The higher the cost of raising the fine relative to the cost of raising the perceived probability, the less steeply sloped the enforcement cost curve. The reason underlying this is illustrated in Exhibit 4-10 by the curves labeled E^1C^1

¹³ The EC curves are therefore "iso-enforcement cost" curves. Similarly, the curve labeled EF is an "iso-expected fine" curve.

Exhibit 4-9

Optimal Values of the Fine and Its Perceived Probability

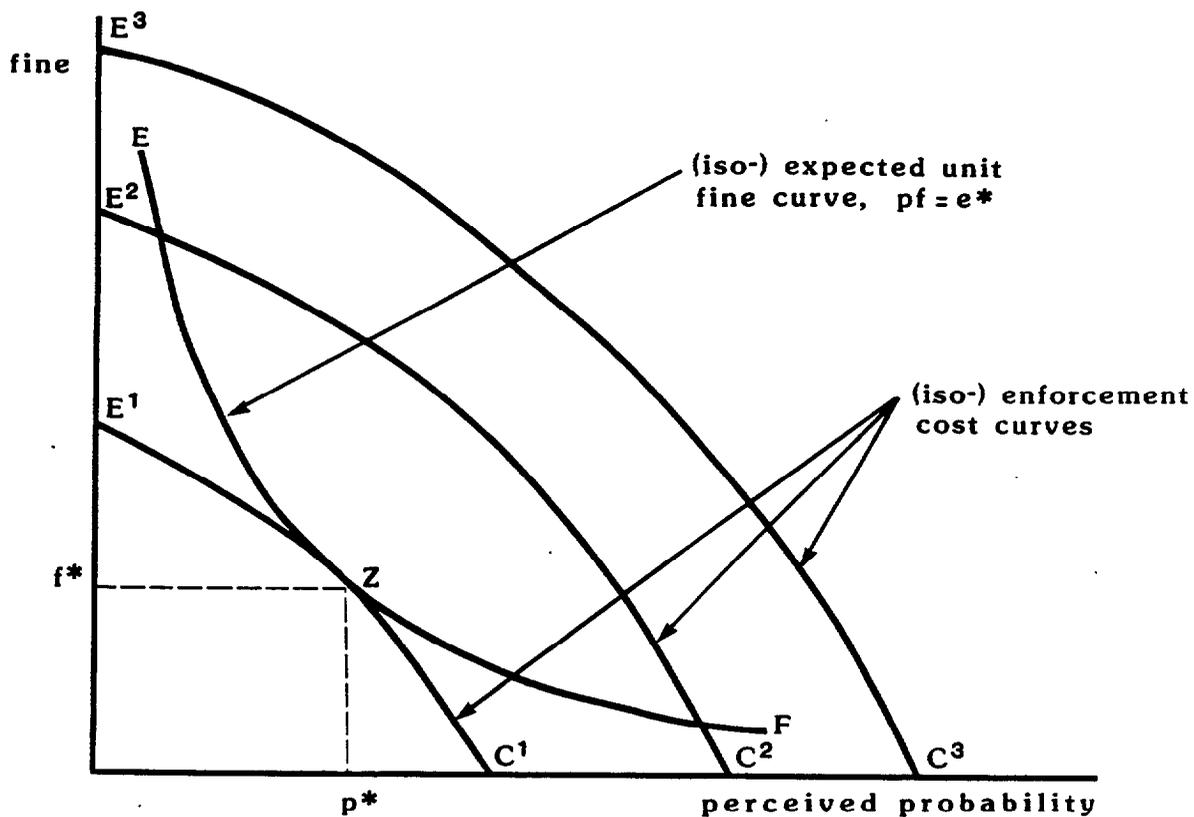
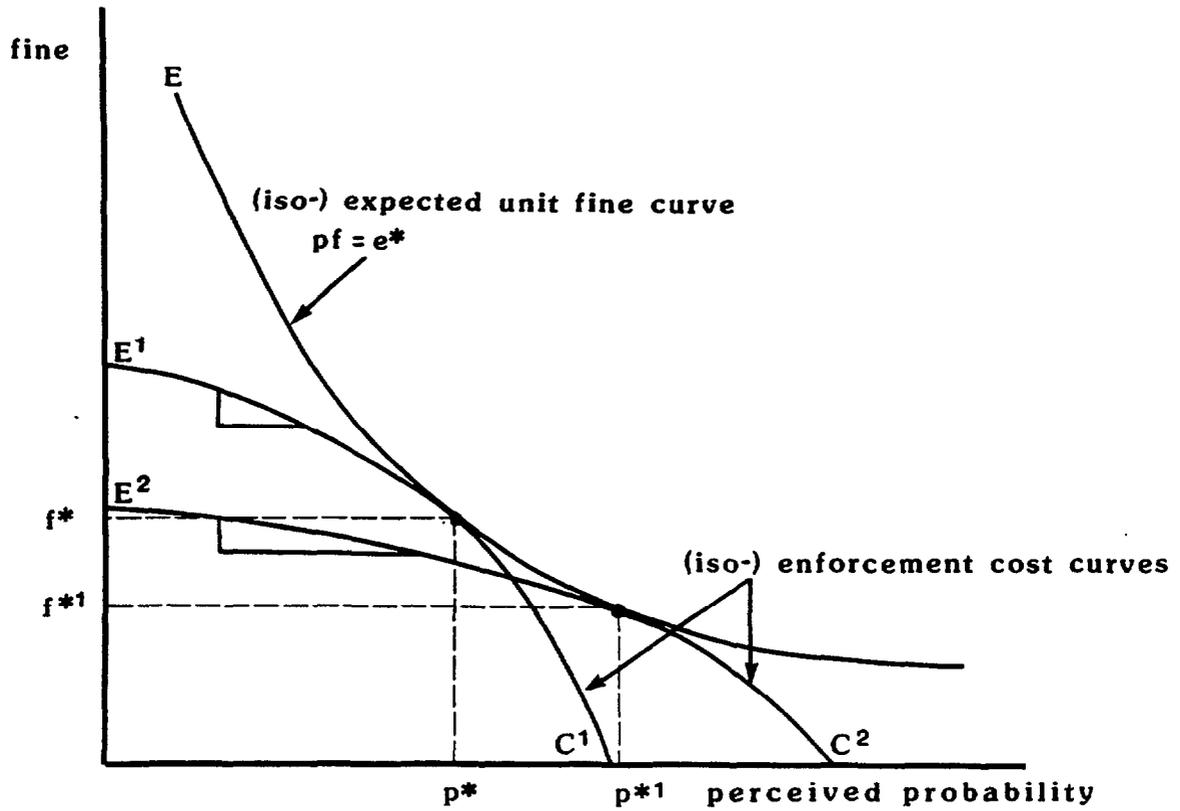


Exhibit 4-10
Relative Costs and the Optimal Fine and Its Probability



and E^2C^2 . The latter curve reflects higher relative costs of increasing the fine, and it is less steeply sloped than E^1C^1 . The flatter slope implies that for a given increase in the fine, a greater reduction in the perceived probability is required if costs are to remain constant.

As shown in Exhibit 4-10, the effect of increasing the relative cost of the fine is to lower the value of the optimal fine from f^* to f^{*1} and raise the value of the optimal perceived probability from p^* to p^{*1} . This result is consistent with intuition. For example, if the lower relative cost of increasing the perceived probability is due to the fact that the set of firms under consideration believe that marginal enforcement activities vastly increase the probability of detection and penalization, then it is socially cheaper to establish a given level of the expected fine by leaning more heavily on raising the perceived probability, rather than by trying to raise the fines themselves.

The above analysis establishes that, for a given expected fine, the precise value of the optimal fine depends on the relative costs of raising the fine and raising the perceived probability. However, the value of the optimal fine also depends on the value of the optimal expected fine. In general, we can expect the value of the optimal fine to increase as the optimal expected fine **increases**.¹⁴ This is shown in Exhibit 4-11. The expected fine corresponding to the curve labeled E^1F^1 is e^{*1} , which is higher than e^* , the expected fine corresponding to the curve labeled EF. As shown, the optimal fine given an expected fine of e^{*1} is higher than the optimal fine given an expected fine of e^* (f^{*1} is greater than f^*).

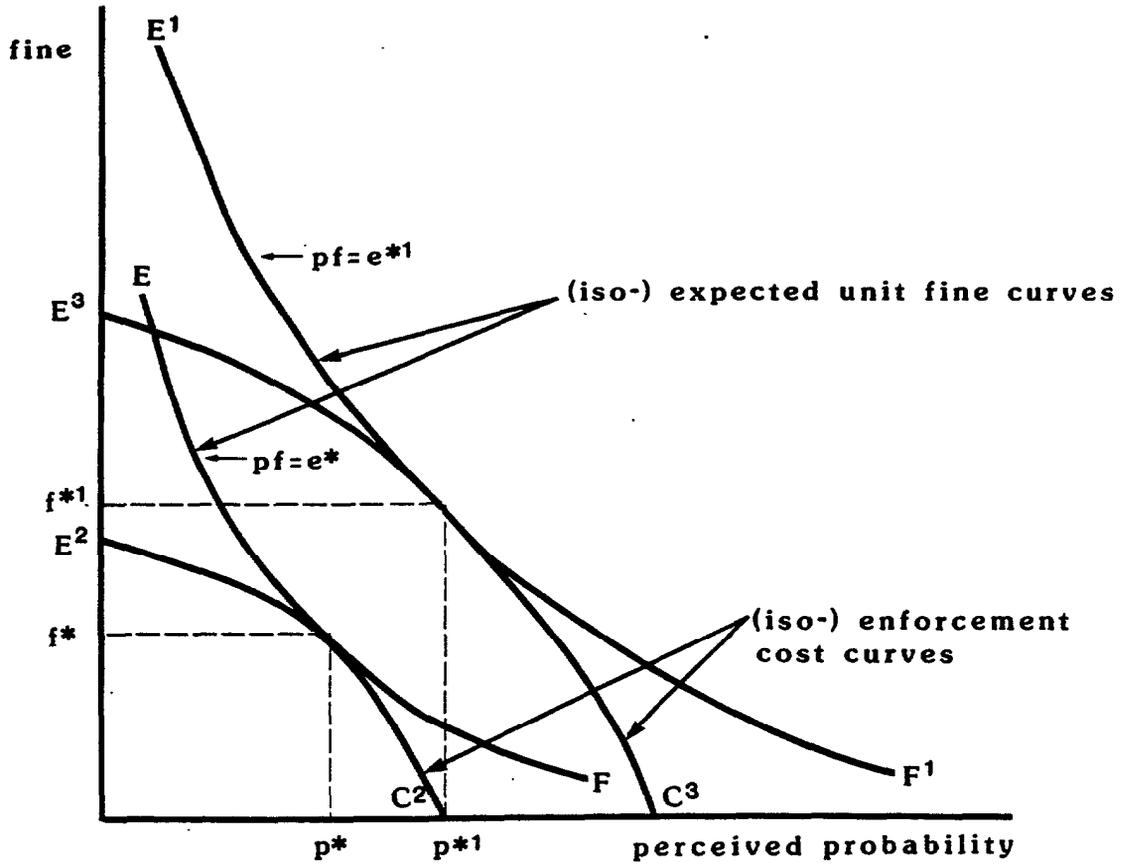
Because the optimal fine is one component of the optimal expected fine, the optimal fine indirectly depends on the same factors as the optimal expected fine, namely: marginal enforcement costs, marginal compliance costs, the relationship between underlying enforcement activities and firms' perceptions of the probability of detection and penalization, and marginal damages avoided. As is true for the optimal expected fine, the optimal fine is a complicated function of these factors. In addition, the direction of the relationship between the optimal fine and each of the three factors is the same as that for the optimal expected fine. For example, higher marginal enforcement costs imply both a lower optimal expected fine and a lower optimal fine.

Thus, the main results of the analysis of the relationship the optimum expected fine and its components -- the optimum fine and its perceived probability -- are that:

¹⁴ It is conceivable in theory for the optimal fine to fall as the optimal expected fine increases, however it is unlikely for this to hold true in reality.

Exhibit 4-11

Relationship Between the Optimal Fine and the Optimal Expected Fine



- The conclusions regarding the relationship between the optimum expected fine and marginal enforcement costs, marginal compliance costs, the perceptions associated with the probability of detection and penalization, and marginal damages avoided, also hold for the optimum fine; and
- The optimal values of the fine and its probability for a fixed expected fine depend on the relative costs of raising the fine and raising the perceived probability. In particular, the higher the cost of raising the fine relative to the cost of increasing the perceived probability, the lower the value of the optimal fine and the greater the perceived probability of being caught.

4.3 INCORPORATING SELF-MONITORING/REPORTING REQUIREMENTS

The model of optimal enforcement presented in the previous section implicitly assumes that violations are detected by random on-site inspections. The model does not incorporate self-monitoring/reporting requirements and the possibility that firms may choose not to report violations. Incorporating these requirements would have complicated the model considerably without materially affecting any of the results. However, it is important to examine the problem of noncompliance with reporting requirements and this is done here by extending the model of the noncompliant firm. The analysis reveals the importance of properly structuring the penalties for violating an effluent limit and for failing to report an effluent limit violation. In particular, it shows how setting the penalties for the two types of penalties independently may present the firm with an incentive to conceal violations.

As noted in Section 2.2.1, existing CWA regulations require that firms monitor their discharges and report any significant violations to the relevant state or federal **agency**.¹⁵ Moreover, discharge data must be periodically submitted to the Agency even when the firm is in compliance. The intent of these regulations is to reduce the burden on state and federal agencies for monitoring discharges. Indeed, if firms complied fully with self-monitoring/reporting requirements, and correctly notified the appropriate agency of any and all violations, there would be no need for EPA or state agencies to conduct on-site inspections. However, in practice, firms may choose to conceal violations either by falsifying their reports or by simply ignoring reporting requirements. As a result, on-site inspections are necessary to provide firms with an incentive to correctly report violations.

¹⁵ We shall examine the relationship between the firm's reporting decision and its decision to exceed the effluent limit further below.

4.3.1 The Firm's Reporting Decision

Exhibit 4-12 presents the "decision tree" of a firm that has chosen to exceed an effluent limit by an amount v (the violation size) and must now decide whether or not to report the violation.¹⁶ By this we mean that the firm decides whether to file its report with information indicating the violation or decides instead not to file the required reports or reports incorrect information. In what follows, "not reporting" means failing to report as required.¹⁷

As shown in Exhibit 4-12, the firm has two options: the first is to report the effluent limit violation and the second is not to report it. If the firm adopts the first option, it is automatically assessed a fine of f per unit violation, and pays a total fine of fv (f times v). If the firm adopts the second option and decides not to report the effluent limit violation, one of two events may occur. The firm could be caught and assessed a fine for failing to report the violation, in addition to paying the fine for the effluent limit violation itself (fv). The fine for failing to report the violation could be either a fixed amount, independent of the magnitude of the violation, or it could be a variable amount that increases with the magnitude of the violation. The latter possibility seems more plausible and we assume that the fine for not reporting a violation is an amount g per unit violation. Therefore, if the firm is caught not reporting a violation, it is assessed a total fine of $f+g$ per unit violation. The expected fine per unit violation is $p(f+g)$, where p denotes the perceived probability that the firm is caught; the total expected fine is equal to the expected fine per unit violation times the violation size: $p(f+g)v$.

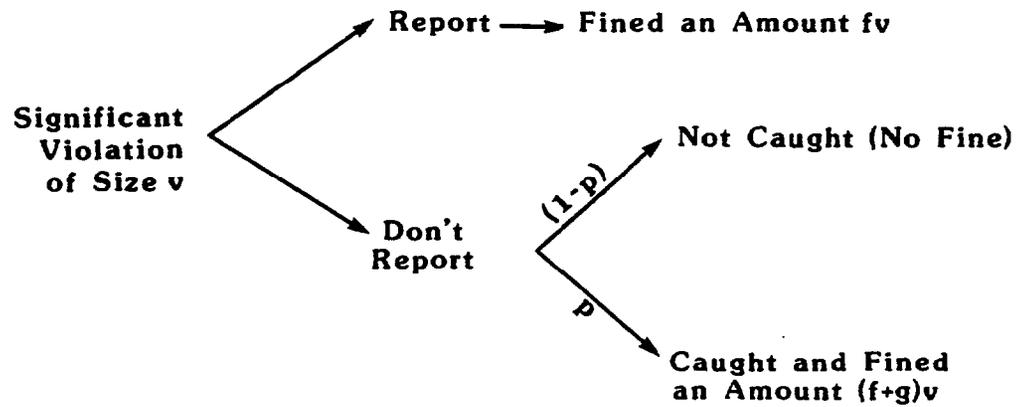
The second possible event, which has a perceived probability of $(1-p)$ of occurring, is that the firm escapes detection and is not caught and fined. For this event, the expected fine is zero since no fine is assessed.

Since we assume the firm's objective is to minimize the total expected costs it incurs, the firm will adopt the reporting option with the lower expected penalty. As shown in Exhibit 4-12, the expected penalty associated

¹⁶ It is assumed that the violation is a significant one and must therefore be reported. The precise definition of "significant" varies from pollutant to pollutant, see Section 2.2.1

¹⁷ We assume that firms honestly report their violations. We could extend the decision tree and incorporate the possibility that firms understate their violations and are penalized for doing so. However, this would complicate the analysis considerably without providing much in the way of additional insight. Also, existing data suggests that false reporting is not a major problem given the severity of the penalties for doing so. Falsifying reports is likely to be viewed as a criminal offense punishable by fines and imprisonment, whereas failure to report a violation is likely to be treated as a civil offense with lower attendant penalties.

Exhibit 4-12
Firm's Reporting Decision Tree



Total Expected Fine if Report = fv

Total Expected Fine if Don't Report = $p(f+g)v$

with the first option (reporting the violation) is simply fv , whereas the expected penalty associated with the second option (not reporting the violation) is $p(f+g)v$. Therefore, the firm will report the violation if the total expected fine associated with reporting the violation is lower than the total expected fine associated with not reporting the violation:

$$fv \leq p(f+g)v, \quad (\text{report violation}) \quad (1)$$

and it will conceal the violation if the opposite is true:

$$fv \geq p(f+g)v \quad (\text{conceal violation}) \quad (2)$$

If it so happens that the expected penalties associated with the two options are identical (i.e., $fv = p(f+g)v$) then the firm will be indifferent between reporting the violation and not reporting it. Examining the above inequalities it is clear the relative magnitude of the expected penalties for the two options depends on the probability that the firm is caught and fined (p), the fine for exceeding the effluent limit (f), and the fine for failing to report a violation (g).

The inequality in equation (1) can be simplified by dividing both sides of the inequality by v , this gives

$$f \leq p(f+g) \quad (\text{report violation}) \quad (3)$$

Multiplying out the right-hand side of this inequality, and subtracting the term pf from both sides, the inequality can be written as

$$(1-p)f \leq pg \quad (\text{report violation})$$

This expression can be rearranged to give

$$g/f \geq (1-p)/p \quad (\text{report violation}) \quad (4)$$

Thus, the firm will choose to report the violation if the ratio of the fine for failing to report the violation to the fine for the violation itself (g/f) is greater than the ratio of the perceived probability of not being caught to the perceived probability of being caught ($(1-p)/p$). This implies that the fine for not reporting a violation must be relatively large for the firm to report a violation because the probability of being caught is likely to be very small. This is demonstrated in the table below where the value of $(1-p)/p$ is calculated for various values of p :

<u>p</u>	<u>(1-p)/p</u>
0.001	999
0.01	99
0.1	9
1.0	0

A plausible value for p is 0.01: in the case of an effluent limit that restricts daily discharges, it implies that the firm anticipates inspection three to four times a year ($0.01 \times 365 \text{ days/year} = 3.65 \text{ days/year}$), which is a common inspection frequency. For a value of $p = 0.01$, the above table indicates that the unit fine for failing to report an effluent limit violation must be 99 times higher than the unit fine for the effluent limit violation itself if firms are to choose the reporting option. Although we do not have data on penalties for failing to report a violation, it appears unlikely that existing penalties are structured such that it is in the firm's best interest to report violations, especially given the frequency with which firms fail to report violations (see Chapter 1). Note, however, that this analysis concerns a failure to report, rather than reporting false information, the latter of which actions could result in criminal penalties and appears to be less prevalent.

Examining equation (3), it can be verified that raising the perceived probability of catching and fining violators (p) or raising the fine for failing to report violations (g) will increase a firm's incentive to report violations. Both these increases raise the expected unit fine associated with failing to report a violation ($p(f+g)$), without changing the expected unit fine associated with reporting a violation (f).

On the other hand, raising the fine for exceeding an effluent limit (f), while holding everything else constant, reduces the incentive for firms to report violations. This can be seen most easily with the aid of equation (4). Increasing the value of f makes the left-hand side of the inequality (g/f) smaller, making it less likely for the inequality to be satisfied. For example, suppose $g = \$1000$, $p = 0.1$, and, initially, $f = \$50$, since

$$g/f = \$1000/\$50 = 20 \leq 9 = (1-0.1)/0.1 = (1-p)/p$$

the firm will report violations. Now suppose that the fine for exceeding the effluent limit is raised from \$50 to \$100. The ratio g/f falls from 20 to 10, but it is still larger than nine, the ratio of $(1-p)$ to p , therefore it is still in the firm's interest to report violations. Now consider a further increase in the fine for exceeding the effluent limit from \$100 to \$150. The ratio g/f falls to 6.67, which is smaller than nine, the ratio of $(1-p)$ to p , therefore, the firm will now choose not to report violations. This result, though surprising, is easily explained: the higher fine for exceeding an effluent limit implies that firms have more to lose if they report their violations, therefore, if the probability of being caught and the fine for not reporting remain the same, the firm has a greater incentive to conceal violations.

This result implies that raising the fine for violations, without simultaneously adjusting the fine for not reporting and/or the perceived probability that firms are caught and fined, could have the undesirable effect of inducing firms that previously reported violations to cease doing so.

4.3.2 Relationship between the Firm's Reporting Decision and its Violation Size

Thus far, the model has simply taken the firm's violation size as given and examined the firm's behavior assuming that it is noncompliant. To complete our analysis, we need to relate the firm's reporting decision to its decision on the extent of noncompliance. Exhibit 4-13(a) presents the firm's marginal compliance costs and the expected unit fine it faces when it reports violations ($EFR = f$) and when it does not report violations ($EFNR = p(f+g)$). Since the expected fine associated with reporting violations is the lower of the two expected fines, the firm will choose to report violations. The relevant expected fine schedule, therefore, is EFR , and the firm's violation size is given by the intersection of EFR and MCC^0 .

If, instead, the figure was drawn such that the expected fine associated with not reporting violations ($EFNR$) were lower than the expected fine associated with reporting violations (EFR), the firm would choose not to report violations, and the relevant expected fine schedule would be $EFNR$. This case is illustrated in Exhibit 4-13(b). The firm's violation size in this case is given by the intersection of $EFNR$ and MCC^0 .

Let us now examine the effect on the firm's violation size of raising the expected fine associated with reporting violations (EFR). Since $EFR = f$, this amounts to increasing the value of the fine for exceeding the effluent limit, shifting the expected fine curve up from EFR to EFR^1 in Exhibit 4-14. The increase in f also shifts up the expected fine associated with not reporting violations since $EFNR = p(f^1+g)$. If the increase in f is relatively small, the expected fine associated with reporting violations will still be lower than the expected fine associated with not reporting violations. As a result, the firm will continue to report violations. This case is shown in Exhibit 4-14(a). Note that the increase in the fine for exceeding the effluent limit (f) lowers the firm's optimal violation size from v^{*0} to v^{*1} .

In contrast, if the increase in f is relatively large, the expected fine associated with reporting violations will exceed the expected fine associated with not reporting violations (recall the numerical example provided earlier). As a result, the firm will switch to not reporting violations. This case is shown in Exhibit 4-14(b). The firm's initial violation is given by the intersection of EFR and MCC^0 . Its violation size after f is raised is given by the intersection of $EFNR$ and MCC^0 . Thus, the increase in the fine for exceeding the effluent limit induces the firm to switch from reporting its violations to concealing its violations, however, it reduces the firm's violation size from v^{*0} to v^{*1} . This result holds true in general: an increase in the fine for exceeding an effluent limit may induce firm's to stop reporting their violations, but it will always reduce the violation size chosen by the firm. This is true because increasing the fine for exceeding the effluent limit unambiguously raises both the expected fine

Exhibit 4-13

Firm's Violation Size and Its Reporting Decision

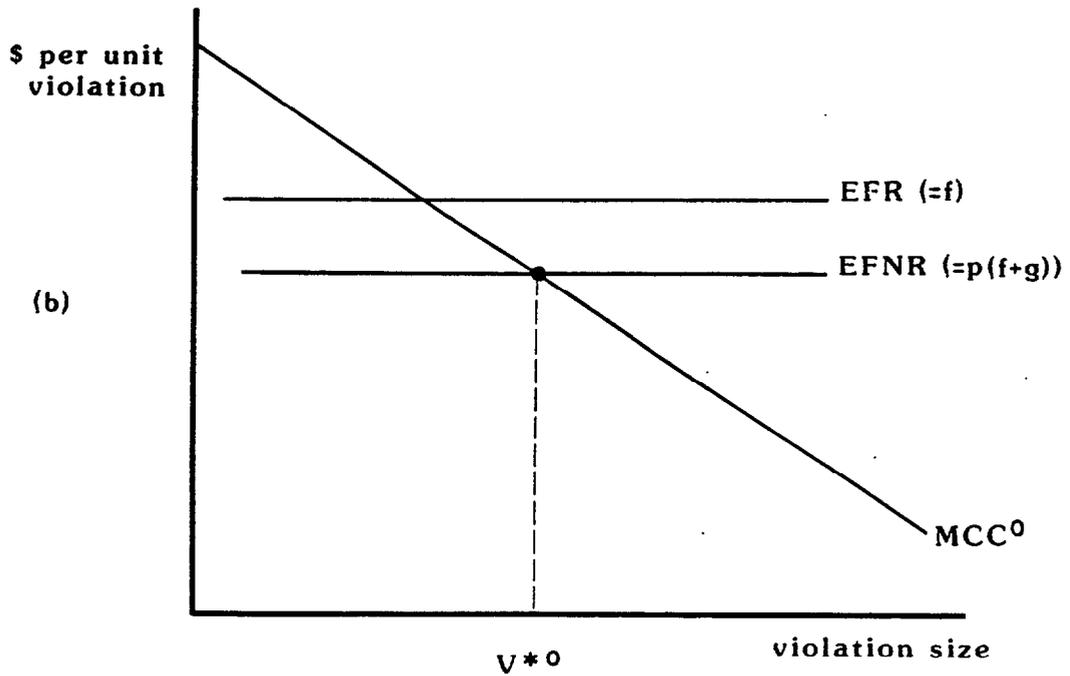
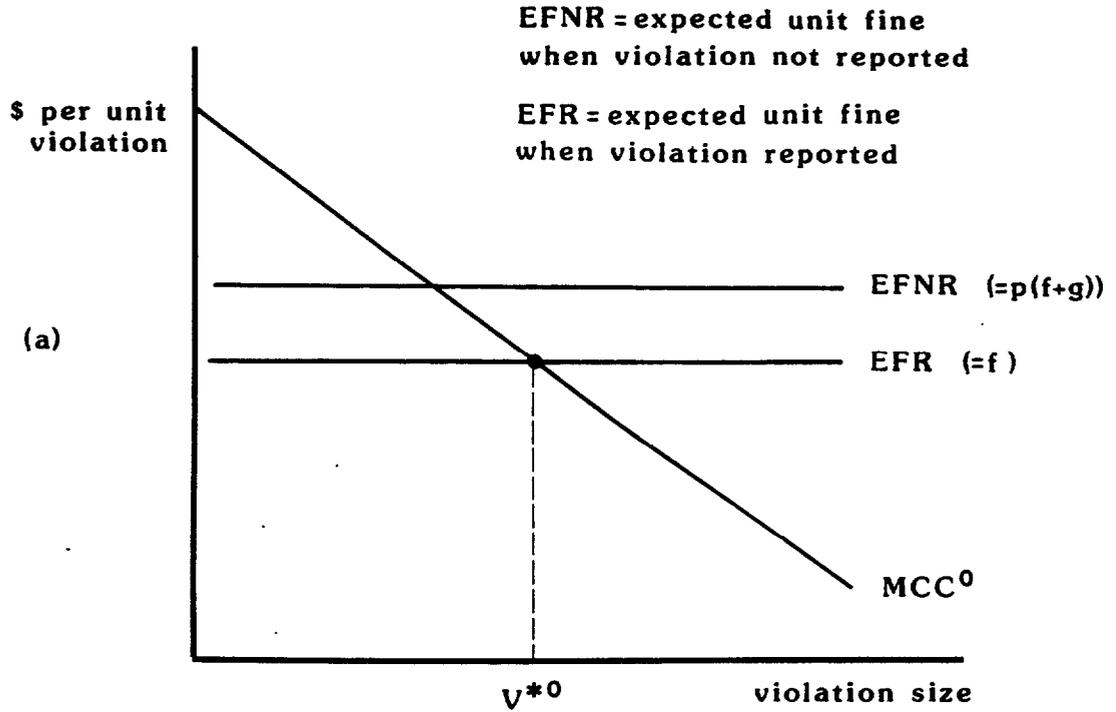
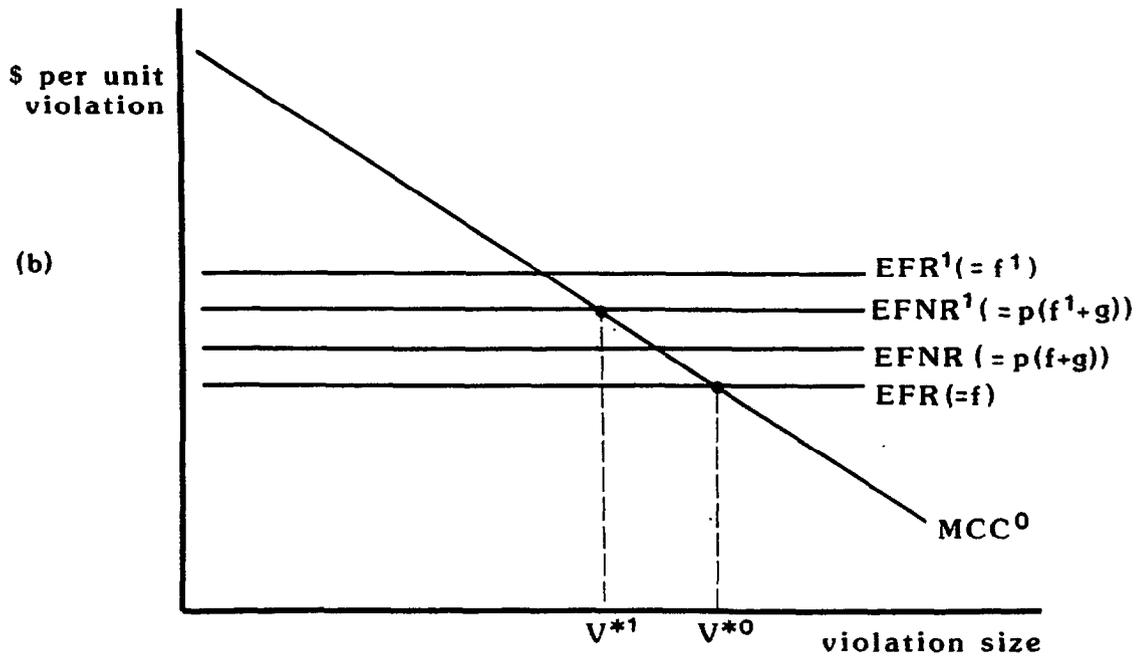
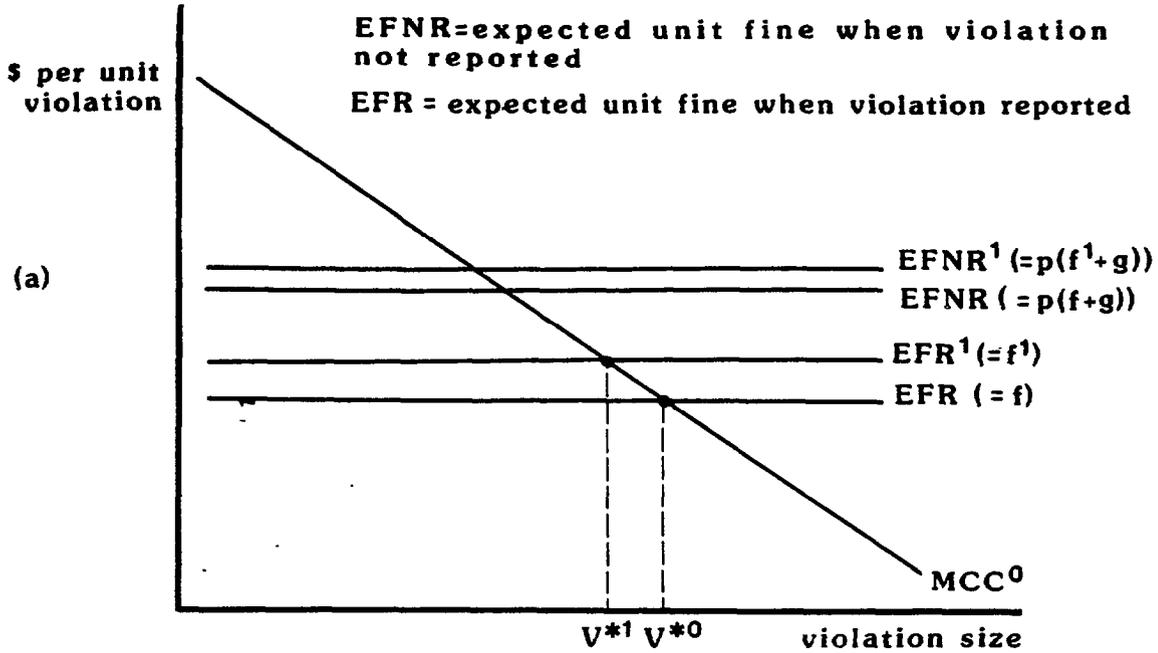


Exhibit 4-14

Effects of an Increase in the Fine for Exceeding an Effluent Limit



associated with reporting a violation and the expected fine associated with not reporting a violation (EFR and EFNR) even though it may change their relative **magnitudes**.¹⁸

We can summarize the results of the analysis of self-monitoring/reporting requirements as follows.

- If the perceived probability that violations are caught and fined is low, the fine for failing to report an effluent limit violation must be substantially larger than the fine for the effluent limit violation itself if firms are to report violations.
- Firms' incentive to report effluent limit violations increases if the perceived probability of catching and fining violations is raised and/or the fine for not reporting violations is raised.
- Raising the fine for effluent limit violations may induce firms to stop reporting violations. However, raising the fine will always reduce the magnitude of firms' violations.

4.4 IMPLICATIONS OF THE MODEL

As noted earlier, the model of enforcement developed in this chapter is really only relevant to the subset of firms that will not comply with environmental regulations unless presented with financial incentives to do so. Thus, all of our conclusions based on the model are limited to this subset of firms. Furthermore, our model of enforcement, as developed in Section 4.2, does not provide a simple method for calculating the optimal expected fine or the optimal fine. Indeed, it is unlikely that any conceptually sound model of enforcement would yield a simple penalty formula. However, the model presented does offer several insights into the characteristics of optimal enforcement. These are presented and discussed below.

4.4.1 Setting Penalties Equal to the Benefits of Noncompliance

As discussed in Chapter 2, EPA penalty policy has emphasized penalties to recoup the benefits to the firm from noncompliance. The alleged rationale for this policy is that it removes the benefits of noncompliance. However, it can

¹⁸ It is assumed above that the firm's violation is a significant one and must therefore be reported. The analysis could easily be modified to take into account the possibility that the firm's violation is small enough so as not to be considered significant. This modification would not affect the results derived.

be shown using the model of the noncompliant firm developed in Section 4.2.1, that this penalty policy deters noncompliance only if the firms in question believe that the probability of catching and fining violators is equal to one, that is, only if firms believe that violations are always detected and penalized.

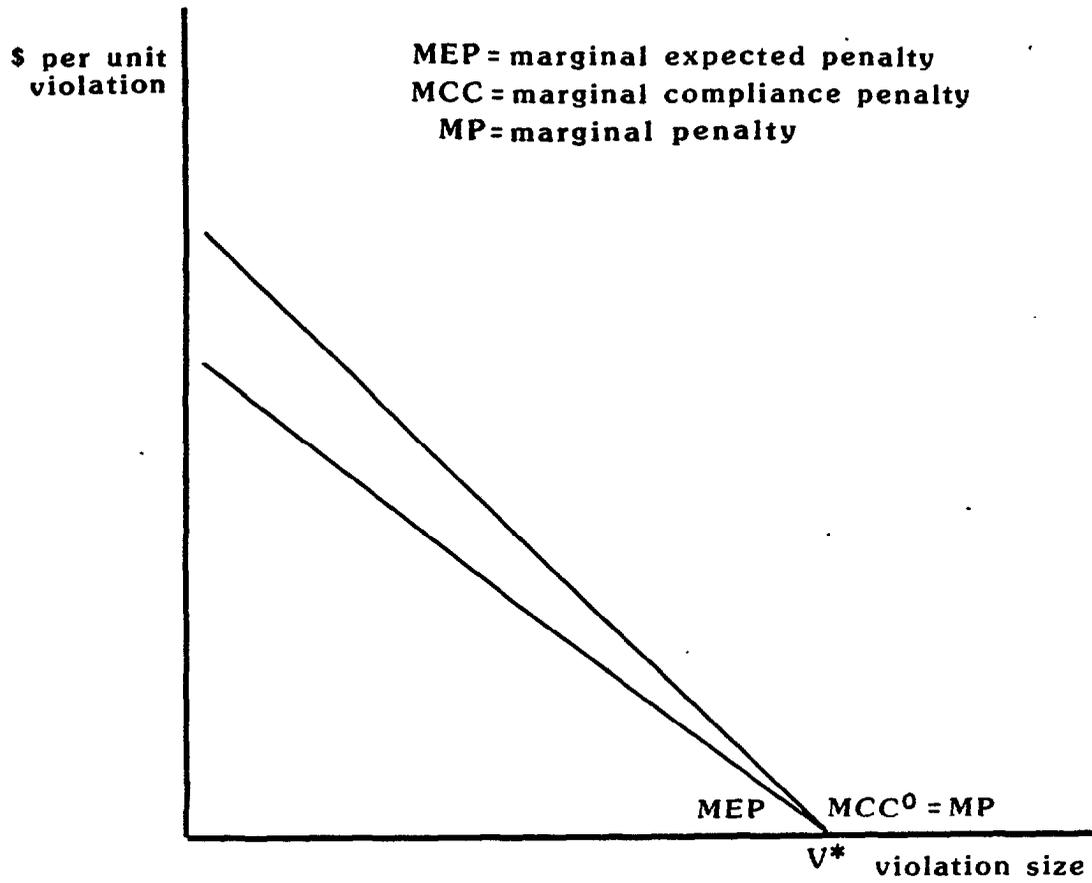
In the context of our model of the noncompliant firm, the benefits of noncompliance are simply the compliance costs avoided. Therefore, the total penalty is simply equal to the compliance costs avoided, and the marginal penalty is equal to the marginal compliance cost ($MP = MCC^0$).¹⁹ Since the latter diminishes as the violation gets larger because initial units of compliance activities are cheaper than those on the margin (see Exhibit 4-1), so does the marginal penalty -- because the marginal penalty is simply the amount by which the penalty changes as compliance changes if the penalty is simply equal to to benefits of noncompliance. (This implies that the fine per unit violation is not constant, contrary to what was assumed in developing the model.) If the perceived probability (p) of catching and fining violators is equal to one, the marginal expected penalty (MEP) is equal to the marginal penalty (MP) since $MEP = p \times MP$. In this case the marginal expected penalty curve and the marginal compliance cost curve are identical. As a result, the firm is indifferent between being in compliance and not being in compliance. Simply stated, if a firm believes that penalties consistently recoup the benefits of noncompliance, the firm will not care whether it is in compliance. However, if the penalty exceeds the benefits of noncompliance by even a small amount, the firm will choose to be perfectly compliant. This may be why the penalty levied is usually augmented by the damages due to noncompliance and/or extraordinary enforcement costs.

In reality, both the objective and the perceived probabilities of catching and fining violations are less than one, that is, some violations go undetected or unpunished and firms know this, although their expectations do not necessarily have to be identical with reality. In this case, the marginal expected penalty lies below the marginal compliance cost curve since $MEP = p \times MP = p \times MCC^0$. This is illustrated in Exhibit 4-15. Since marginal compliance costs now exceed the marginal expected penalty for all violation sizes, the firm sets its violation equal to v^* , the violation level at which it incurs no costs on compliance. Thus, setting the penalty equal to the benefits from noncompliance does not promote compliance if firms believe that the probability of catching and fining violators is less than one. Of course, if the penalty is augmented by other factors, such as the damages from noncompliance, the firm may choose to be compliant, but this will depend on the precise amount by which the penalty is augmented.

¹⁹ More precisely, the marginal penalty is equal to the negative of the marginal compliance cost, however, this does not affect our exposition our conclusions.

Exhibit 4-15

Violation Size When Penalties Are Equal to the Benefits of Noncompliance



Clearly, the perceived probability of catching and fining violators is critical to the success of a penalty policy based on the benefits from noncompliance. In practice, the magnitude of the perceived probability will depend on the type of violation being considered, Agency enforcement activities, and subjective factors, such as publicity. For day-to-day violations of effluent limits, it is highly unlikely that firms believe that each and every violation will be detected and punished. For such violations the relevant perceived probability is likely to be very small unless the violation has an acute and easily observed effect on environmental quality. For other violations, such as failure to install equipment, the perceived probability of being caught and fined may well equal one because if the firm is inspected at all, it is relatively easy to determine whether or not the firm is in compliance and how long it has not been in compliance (i.e., the period over which the firm failed to install the necessary equipment). Thus, a benefits-based enforcement policy (i.e., one that targets resources and levies penalties based on damages) may be effective in deterring certain types of violations but not others.

4.4.2 The Optimum Penalty for Effluent Limit Violations

As discussed above, the model of the noncompliant firm reveals that setting penalties equal to the benefits of noncompliance may not deter noncompliance. The model of optimal enforcement further implies that such a penalty, or even a penalty set equal to the sum of the benefits from noncompliance, the damages from noncompliance, and the costs of enforcement, is unlikely to be the optimal penalty. Although the model indicates that the optimal penalty is a function of the four factors listed, it depends on them in a fairly complicated way. Furthermore, the optimal penalty is only one of the two key components of an optimal expected penalty, the other being enforcement strategy (i.e., setting the perceived probability). Unfortunately, any general conclusions regarding a formula for the optimal penalty and the optimal expected penalty cannot be provided since it depends on the precise characteristics of the compliance and enforcement costs as well as damages.

4.4.3 Penalties for Failing to Report Violations

Our analysis of self-monitoring/reporting requirements shows that the penalty for failing to report an effluent limit violation must be considerably higher than the penalty for the effluent limit violations itself if firms are to have an economic incentive to report violations. The precise amount by which the reporting penalty must exceed the effluent violation penalty depends on the perceptions of firms regarding the probability that they will be caught and penalized for violations. The smaller this perceived probability, the larger the amount by which the reporting penalty must exceed the effluent violation penalty.

The analysis also shows that, if incentives to report violations are to be preserved, any increases in the penalty for effluent violations should be reflected in the penalty for failing to report violations. Otherwise, it is

possible that increases in the penalty for effluent violations., without compensating increases in the penalty for failure to report violations, could induce firms to stop reporting violations.

4.4.4 Targetting Enforcement Resources at High Damage Violators

Among the conclusions drawn in Section 4.2.2 is that the optimal expected fine (and the optimal fine) rises when marginal damages avoided increase, but marginal enforcement and compliance costs remain the same. This result has important implications regarding the targetting of enforcement resources.

As discussed in the introduction to Section 4.2.2, in the context of the model, a higher expected fine is equivalent to a higher level of enforcement, either through higher penalties themselves or increased enforcement activities. The result cited above therefore implies that if the costs of bringing enforcement action against each of two (or more) firms is roughly similar, and the firms have similar compliance costs, then priority should be given to the firm (or firms) that impose higher damages as a result of their noncompliance. Thus, these firms should be subject to both more frequent monitoring, which raises their objective and, presumably, their perceived probability of being caught, and to larger fines when caught (unless there are overriding deterrence considerations).

4.4.5 Targetting Enforcement Resources at Low Enforcement Cost Violators

Another conclusion drawn in Section 4.2.2 is that the optimal expected fine (and the optimal fine) falls when marginal enforcement costs go up, but marginal compliance costs and marginal damages avoided remain the same. This conclusion has implications analagous to those of the previous result discussed. In particular, it implies that if two (or more) firms have similar compliance costs and impose similar damages (in monetary terms) as a result of their noncompliance, then priority should be given to taking enforcement action against violators for which enforcement is less costly, unless there are overriding deterrence considerations.

Similarly, the analysis in Section 4.2.2 also suggests that the optimal expected fine should be higher for firms or activities in which the relationship between the perceived probability of being caught and penalized and the underlying objective enforcement activities is higher. However, in general, the level of enforcement resources required to achieve this higher expected penalty will be less than in other circumstances.

4.5.5 Mix of the Optimal Fine and the Optimal Perceived Probability

Finally, the model also suggests that the optimal settings of the two components of the expected fine are predicated on the relative costs of increasing the expected fine through the two avenues. Thus, to the extent that raising the expected fine through increases in the perceived probability is easier and cheaper than trying to collect higher fines from firms when they are caught, this method should be used more extensively.

APPENDIX A

SAMPLE CALCULATION OF OPTIMAL EXPECTED FINE

This appendix presents an example demonstrating the calculation of the optimal expected fine for an effluent limit violation. This example makes more tangible many of the concepts and results discussed in Section 4.2.

As noted in the body of the text, to calculate the optimal expected fine we need detailed information on the firm's compliance costs, the damages associated with violations, and the costs of enforcement. In the example presented below, this information is assumed by specifying a compliance cost function, a damages avoided function, and an enforcement cost function. These functions are presented and explained below. No particular significance should be attached to the functions used in the example. They were chosen primarily because they yielded a simple formula for the optimal expected fine and because they were consistent with a priori notions about the characteristics they should have. For instance, the compliance cost function should be such that costs fall as the violation size increases.

Compliance Costs

The compliance cost function assumed is

$$CC^0 = 0.5c(v - v_m)^2,$$

the parameter v_m denotes the maximum violation size the firm would choose, and

c is an arbitrary compliance cost parameter. The higher the value of c , the higher are the costs of compliance. A graph of the compliance cost function would look very much like the curve labeled CC^0 in Exhibit 4-1(a). As the violation size, v , increases, compliance costs fall, reaching a minimum of zero when the firm sets its violation size equal to v_m . The firm would never

set its violation size above v_m because this would be costly: the firm would have to devote resources just to generating pollution.

The marginal costs of compliance (obtained by taking the derivative of the cost function with respect to the violation size) are given by

$$MCC^0 = -c(v - v_m).$$

Since the violation size chosen by the firm is always smaller than v_m , and c

is assumed to be a positive number, marginal compliance costs are negative. This just reflects the fact that compliance costs fall as the violation gets larger. A graph of the marginal compliance cost function would resemble the curve labeled MCC^0 in Exhibit 4-1(b), with marginal compliance costs falling as the violation gets larger. (Note, as pointed out earlier, that the negative of marginal compliance costs are actually plotted in Exhibit 4-1(b).)

Damages Avoided

The damages avoided function used in the example is

$$DA^0 = d(v - v_m),$$

where v_m once again denotes the firm's maximum violation size, and d is an arbitrary damages-avoided parameter. The higher the value of d , the larger the value of the damages that are avoided. A graph of the damages avoided function would be virtually identical to the curve labeled DA^0 in Exhibit 4-1(a), with damages avoided falling as the violation size increases, reaching a minimum of zero when the firm's violation size is at its maximum value.

Marginal damages avoided (obtained by taking the derivative the damages avoided function with respect to violation size) are given by

$$MDA^0 = -d.$$

Therefore, they are independent of violation size, as depicted in Exhibit 4-1(b).

(Note that the negative of marginal damages avoided are actually plotted in the figure.)

Enforcement Costs

The enforcement cost function assumed in the example is

$$EC = 0.5he^2,$$

where e is the expected fine, and h is an arbitrary enforcement cost parameter. The higher the value of h , the more expensive it is to raise the expected fine. Enforcement costs are assumed to increase exponentially with the expected fine, as depicted by the curve labeled EC in Exhibit 4-4. Furthermore, this particular function assumes a given relationship between objective enforcement activities and firms' perceived probability of detection and penalization.

Marginal enforcement costs (obtained by taking the derivative of the enforcement cost function with respect to the expected fine) are given by

$$MEC = he.$$

Marginal enforcement costs therefore rise linearly as the expected fine goes up, as depicted by the curve labeled MEC in Exhibit 4-5.

The information presented above allows us to calculate the optimal expected fine as a function of the parameters c , d , and h . Given estimates of these three parameters, we could compute the numerical value of the optimal expected fine.

The Firm's Violation Size

Following the procedure outlined in Section 4.2, the first step in calculating the optimal expected fine is determining the relationship between the expected fine and the firm's violation size. As explained in Section 4.2.1, the firm's violation size is the violation size at which the (negative) marginal costs of compliance equal the expected fine per unit violation (e). In terms of the marginal compliance cost function presented above, it is the violation size at which:

$$-MCC^0 = c(v_m - v) = e = EF$$

By solving this equation for v , the violation size, v^* , is obtained:

$$v^* = v_m - e/c.$$

Thus, the violation size is equal to the maximum violation size the firm would choose minus the expected fine divided by the compliance cost parameter.

Examining the expression for v^* , it is clear that the firm's violation size declines as the expected fine goes up, which is consistent with the general result derived graphically in Section 4.2.1. At one extreme, if the expected fine is equal to zero, the firm's violation size is equal to its maximum violation size. At the other extreme, if the ratio of the expected fine to the compliance cost parameter (e/c) is greater than or equal to the maximum violation size (v_m), the firm's violation size is zero. (Contrary to

what the expression for v^* indicates, the firm would not choose a negative violation size because this would imply that it is incurring expenses to restrict discharges below the allowed level. As such, it would not be minimizing costs.)

Examining the expression for v^* it is also clear that the firm's violation size increases as compliance costs go up: the larger the value of the compliance cost parameter, c , the smaller is e/c , and the larger is the violation size. Once again, this is consistent with the general result derived graphically in Section 4.2.1.

Deriving Compliance Costs and Damages Avoided as a Function of the Expected Fine

Given an expression for the firm's violation size as a function of the expected fine, the second step in calculating the optimal expected fine is deriving expressions for compliance costs and damages avoided as a function of the expected fine. (Note that enforcement costs are naturally a function of the expected fine.) All this entails is substituting the expression for the violation size ($v^* = v_m - e/c$) for v in the compliance cost and damages

avoided function, CC^0 and DA^0 . In the case of the compliance cost function, this gives:

$$\begin{aligned} CC &= 0.5c(v_m - v + e/c)^2 \\ &= 0.5c(e/c)^2 = 0.5c(e^2/c^2) \\ &= 0.5e^2/c. \end{aligned}$$

And in the case of the damages avoided function, this yields:

$$\begin{aligned} DA &= d(v_m - v_m + e/c) \\ &= de/c. \end{aligned}$$

For any particular value of the expected fine, e , these two functions give the firm's compliance costs and the damages avoided by automatically taking into account the relationship between the expected fine and the violation size the firm would choose.

Deriving Marginal Compliance Costs and Damages Avoided as a Function of the Expected Fine

The penultimate step in calculating the optimal expected fine is deriving expressions for marginal compliance costs and marginal damages avoided as a function of the expected fine. Marginal compliance costs as a function of the expected fine are given by

$$MCC = e/c$$

(derived by taking the derivative of the function CC with respect to the expected fine). Therefore, marginal compliance costs as a function of the expected fine rise linearly with the expected fine, as depicted by the curve labeled MCC in Exhibit 4-5.

Marginal damages avoided as a function of the expected fine are given by

$$MDA = d/c$$

(obtained by taking the derivative of the function DA with respect to the expected fine). Thus, marginal damages avoided are independent of the expected fine, as depicted by the curve labeled MDA in Exhibit 4-5.

As was seen above, marginal enforcement costs (MEC) rise linearly with the expected fine, and are given by

$$MEC = he,$$

where h is the enforcement cost parameter.

The Optimal Expected Fine

As explained in Section 4.2.2, the optimal expected fine is the value of the expected fine at which the sum of marginal enforcement costs and marginal compliance costs as a function of the expected fine (MEC+MCC) is equal to marginal damages avoided as a function of the expected fine (MDA). Therefore, in the case of our example, it is the value of the expected fine at which:

$$MEC + MCC = he + e/c = d/c = MDA.$$

By solving this equation for e , the optimal expected fine (e^*) is obtained:

$$e^* = d/(1 + hc).$$

Clearly, the optimal expected fine depends on the damages avoided parameter, d , the enforcement cost parameter, h , and the compliance cost parameter, c . Examining the expression for e^* , it can be seen that the optimal expected fine goes up when damages avoided increase (i.e., the value of d increases); and it goes down when enforcement becomes more costly (i.e., the value of h increases). Both these observations are consistent with the general results derived graphically in Section 4.2.2.

The expression for e^* also reveals that for this specific example, the optimal expected fine goes down when compliance becomes more costly (i.e., the value of c increases). As explained in Section 4.2.2., in general, more costly compliance could imply a lower or a higher optimal expected fine. However, in this example, it implies a lower optimal expected fine. Had we used a different set of compliance cost and damages avoided functions, the opposite may have been true.

Given the assumptions on compliance costs, damages avoided, and enforcement costs, the expression for the optimal expected fine is a fairly simple one. It shows that, in general, the optimal total expected fine, which is the optimal expected fine per unit violation times the violation size (e^*v) is not simply the sum of the benefits from noncompliance (compliance costs avoided), the damages from noncompliance (the negative of damages avoided) and enforcement costs.

APPENDIX B

EFFECT OF INCREASED MARGINAL COMPLIANCE COSTS ON OPTIMAL ENFORCEMENT POLICY

This appendix presents an analysis of the ambiguous conclusions concerning optimal enforcement policy of changes in the marginal compliance costs of the firms. This discussion appears in this appendix because of its technical nature.

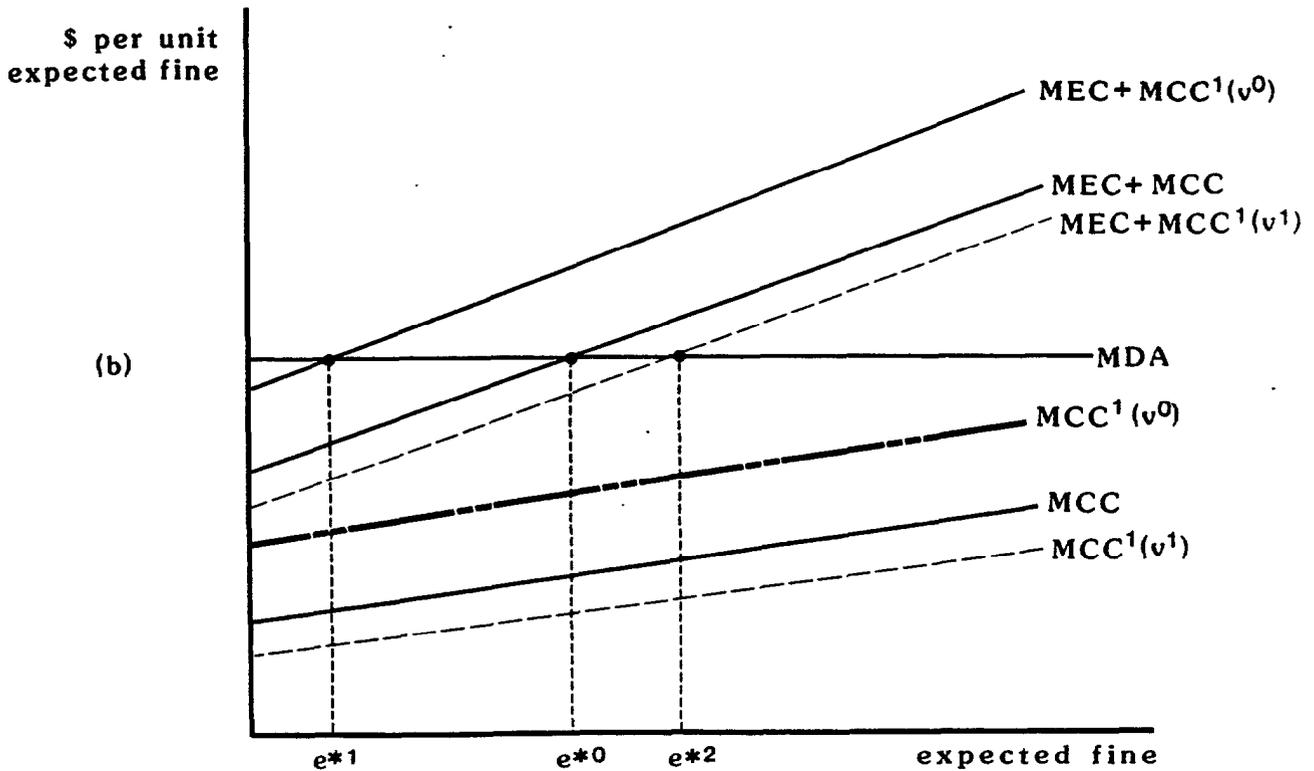
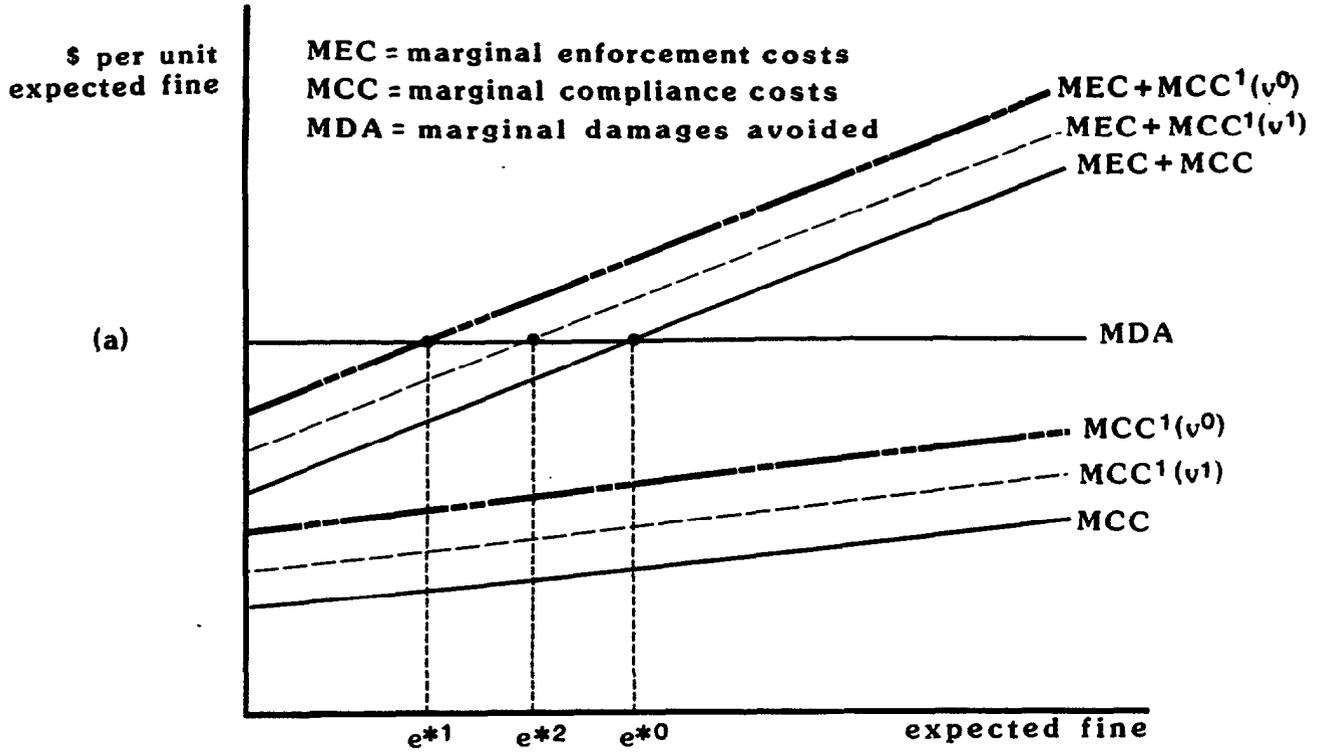
As in the case of marginal damages avoided, the effect of an increase in marginal compliance costs can be expressed as a function of either violation size (MCC^0) or expected fine (MCC). Unlike the case for marginal damages avoided, however, shifts in the marginal costs of compliance when expressed as a function of violation size may result in shifts in the opposite direction for marginal compliance costs expressed as a function of the expected fine. Thus, it is not clear theoretically whether an increase in the marginal compliance costs faced by the firm will tend to raise the optimal fine or to reduce it.

The reason for the ambiguity in the effect of an increase in marginal compliance costs when expressed as a function of violation size can be understood by carefully examining the relationship between the MCC^0 curve and the MCC curve. An upward shift in the MCC^0 curve always shifts up the MCC curve provided we hold the firm's violation constant. This is shown in Exhibit 4-16(a). The initial optimal expected fine is e^{*0} . The upward shift in the MCC^0 curve with the violation held fixed at v^0 (the initial violation size) pushes up the MCC curve to $MCC^1(v^0)$. The MEC+MCC curve is similarly pushed up. The optimal expected fine is now given by the intersection of $MEC+MCC^1(v^0)$ and MDA. Examining Exhibit 4-16(a) it is clear that the new optimal expected fine (e^{*1}) is smaller than the original one (e^{*0}). This change in the optimal expected fine makes sense given the assumption that the firm's violation size is unchanged because the effect of the upward shift in the MCC^0 and MCC curves is to raise the total marginal costs of enforcement without changing the marginal benefits. It follows that fewer resources should be devoted to enforcement relative to the initial allocation.

In reality, however, shifting up the MCC^0 curve will not leave a firm's violation size unchanged. As we established in Section 4.2.1, the effect of such a shift is to raise the firm's violation size (from v^0 to v^1 in Exhibit 4-3). In terms of the MCC curve, this increase in the violation size counters the effect of the upward shift in the MCC^0 curve because a higher violation size implies lower marginal compliance costs along the MCC^0 curve. As a result, the marginal compliance cost curve expressed as a function of the expected fine shifts down from where it is when the violation is held constant at its initial value of v^0 . This is also shown in Exhibit 4-16(a). With the new violation size v^1 , the marginal compliance cost curve shifts down from $MCC^1(v^0)$ to $MCC^1(v^1)$, but it is still above the original marginal compliance cost curve MCC^0 . As a result (for this case at

Exhibit 4-16

Decomposition of the Effects of Increases in Marginal Compliance Costs on the Optimal Expected Fine



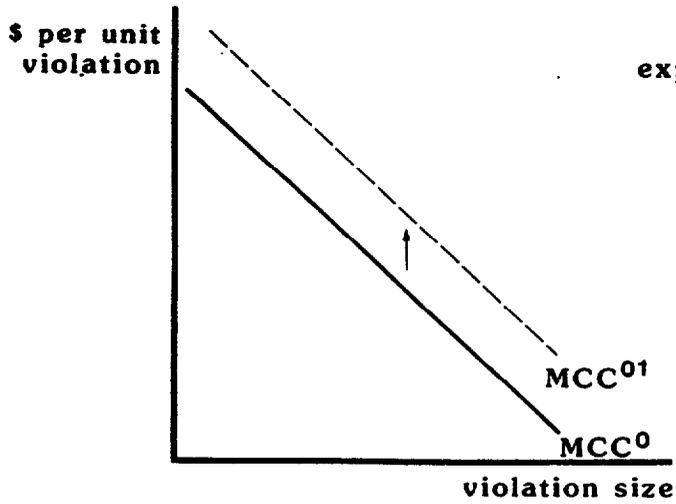
least), the new optimal expected fine, e^{*2} , is smaller than the initial optimal expected fine, e^{*0} , but larger than the optimal expected fine, e^{*1} (if we ignored the firm's response to higher compliance costs by changing its violation size).

On the other hand, the increase in the violation size could be very large, so that it is possible for the marginal compliance cost curve to shift down below the original curve MCC. This is shown in Exhibit 4-16(b). The increase in the violation size from v^0 to v^1 shifts the marginal compliance cost curve down from $MCC^1(v^0)$ to $MCC^1(v^1)$, which is below the original marginal compliance cost curve, MCC^0 . As a result, the new optimal expected fine, e^{*2} , is larger than the initial optimal expected fine, e^{*0} (and, of course, higher than the optimal expected fine, e^{*1} , given that we ignored the firm's response to increased compliance costs by changing its violation size).

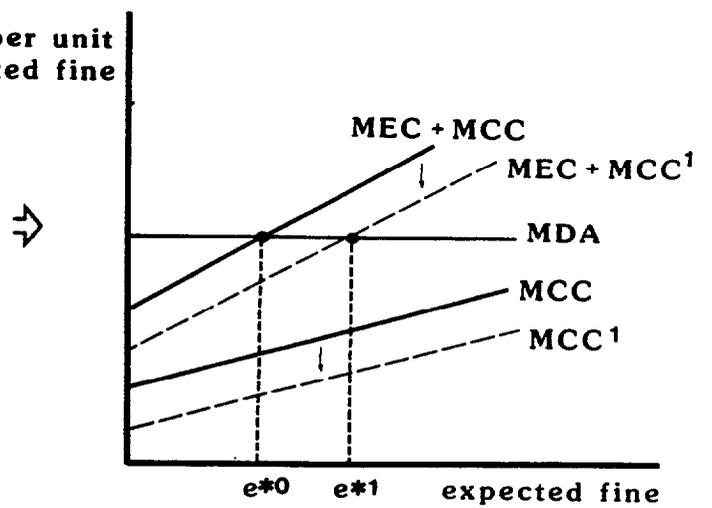
As result of this ambiguity, it is not clear a priori whether an increase in the marginal compliance costs faced by a firm will tend to increase or reduce the optimal expected fine. It is possible for the optimal expected fine to either increase or to decrease in response to increased compliance costs depending on the firms' response. Exhibit 4-17 shows both possible cases. In Exhibit 4-17, the response of the firm to increased marginal compliance costs (the shift of MMC^0 to MCC^{01}) is strong enough to imply that the optimal expected fine should rise from e^{*0} to e^{*1} . Exhibit 4-17(b), however, shows the opposite case in which the optimal expected fine falls when marginal compliance costs rise (MMC^0 shifts upward to MCC^{01}).

Exhibit 4-17

Effects of Higher Marginal Compliance Costs on the Optimal Expected Fine

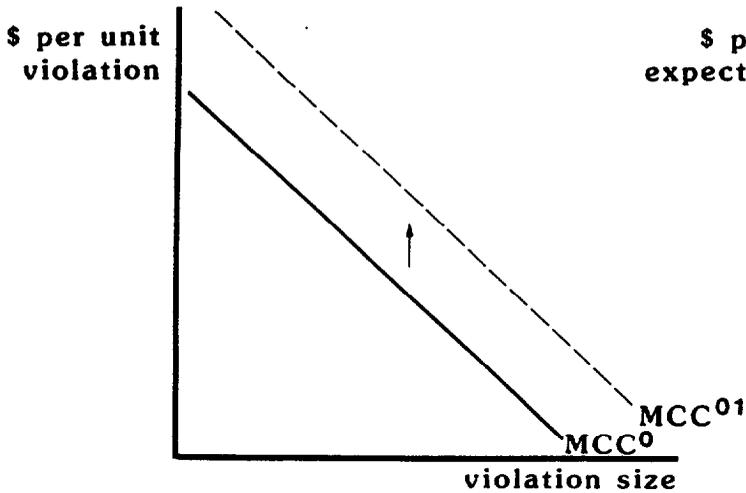


MCC = marginal compliance costs

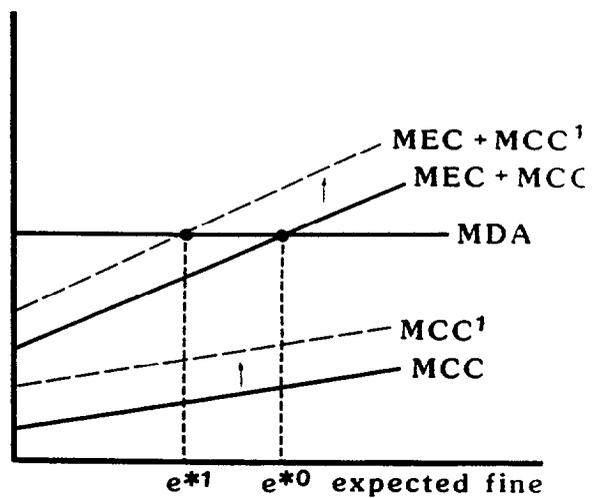


(a)

MEC = marginal enforcement costs
MDA = marginal damages avoided



\$ per unit expected fine



SELECTED REFERENCES

- Beavis, B. and Walker, M. (1981), "Random Wastes, Imperfect Monitoring, and Environmental Quality Standards", Journal of Public Economics, 21:377-387.
- Becker, G. (1968), "Crime and Punishment: An Economic Approach", Journal of Political Economy, 76:169-217.
- Downing, P.B. and Watson, D., Jr. (1974), "The Economics of Enforcing Air Pollution Controls", Journal of Environmental Economics and Management, 1:219-236.
- Environmental Law Institute (1984), Citizen Suits: An Analysis of Citizen Enforcement Actions Under EPA-Administered Statutes, Washington, D.C..
- Federal Register, U.S. Government Printing Office, Vol.50, No. 165, August 26, 1985, Washington, D.C.
- Harford, J.D. (1978), "Firm Behavior Under Imperfectly Enforceable Pollution Standards and Taxes", Journal of Environmental Economics and Management, 5:26-43.
- Heinecke, J.M. (1978), Economic Models of Criminal Behavior, North-Holland Amsterdam.
- Management Advisory Group (MAG) to Federal Construction Grants Program (1985), MAG NPDES Compliance Report, U.S. Environmental Protection Agency, Washington, D.C.
- McKean, R.N. (1980), "Enforcement Costs in Environmental and Safety Regulation", Policy Analysis, 6:269-289.
- Miller, J.G. (1983), "Private Enforcement of Federal Pollution Control Laws", Environmental Law Reporter, 13:10309-10323 (Part I), 14:10063-10082 (Part II), 14:10407-10429.
- Pyle, D.J. (1983), The Economics of Crime and Law Enforcement, St. Martin's, New York.
- Russell, C.S. (1982), Resources for the Future Pollution Monitoring Survey, Summary Report, Resources for the Future, Washington, D.C.
- Russell, C.S., Harrington, W., and W.J. Vaughan (1985), Monitoring and Enforcement in Pollution Control, draft manuscript, Resources for the Future, Washington, D.C.
- Stigler, G.J. (1970), "The Optimum Enforcement of Laws", Journal of Political Economy, 78:526-536.

- Storey, D.J. and McCabe, P.J. (1980), "The Criminal Waste Discharger", Scottish Journal of Political Economy, 27:30-40
- U.S. Environmental Protection Agency, Office of Enforcement and Compliance Monitoring (1984), Agencywide Compliance and Enforcement Strategy and Strategy Framework for EPA Compliance Programs, Washington, D.C.
- U.S. Environmental Protection Agency, Office of Enforcement and Compliance Monitoring (1984), NPDES Program: Use and Level of Penalties, Washington, D.C.
- U.S. Environmental Protection Agency, Office of Policy Planning and Evaluation and Office of Enforcement and Compliance Monitoring (1985), BEN User's Manual, Washington, D.C.
- U.S. Environmental Protection Agency, Office of Water (1985), Clean Water Act Penalty Policy for Civil Settlement Negotiations (Draft Report), Washington, D.C.
- U.S. Environmental Protection Agency, Office of Water (1985), The Enforcement Management System - National Pollutant Discharge Elimination System, Washington, D.C.
- U.S. Environmental Protection Agency, Office of Water and Enforcement Permits (1986), Guidance for Preparation of Quarterly and Semi-Annual Noncompliance Reports, Washington, D.C.
- U.S. Environmental Protection Agency, Office of Water and Enforcement Permits (1979), NPDES Permit Classification Criteria-Current Status, Washington, D.C.
- U.S. General Accounting Office (1983), Wastewater Dischargers Are Not Complying with EPA Pollution Control Permits, GAO/RCED-84-53, Washington, D.C.
- Wasserman, C. (1985), Improving the Efficiency and Effectiveness of Compliance Monitoring and Enforcement of Environmental Policies -- United States: A National Review, unpublished report prepared for the Organization for Economic Cooperation and Development, Environment Directorate, Washington, D.C.

