Marginal Deterrence in Enforcement of Law

Dilip Mookherjee

Indian Statistical Institute

I. P. L. Png

Hong Kong University of Science and Technology and University of California, Los Angeles

We characterize optimal enforcement in a setting in which individuals can select among various levels of some activity, all of which are monitored at the same rate but may be prosecuted and punished at varying rates. For less harmful acts, marginal expected penalties ought to fall short of marginal harms caused. Indeed, some range of very minor acts should be legalized. For more harmful acts, whether marginal expected penalties should fall short of, or exceed, marginal harms depends on the balance between monitoring and prosecution/punishment costs. We also explore how the optimal enforcement policy varies with changes in these costs.

I. Introduction

Laws and regulations to guard the public interest are respected only to the extent that they are enforced. Since all aspects of enforcement—detection, prosecution, and punishment—consume resources, two important issues arise. First, to what extent should society enforce compliance: in particular, what should be the relation between expected penalties on an activity and the harm that it inflicts? Second, how should enforcement policies be adjusted as enforcement costs change? Similar questions confront private parties with contract

We thank the referees, Cyrus Chu, Ed Lazear, Mitch Polinsky, Richard Posner, Eric Rasmusen, Michael Waldman, colleagues from Hong Kong University of Science and Technology and UCLA, and seminar participants at Toronto for encouragement and advice, and B. Ramaswami for help with the graphics.

[*Journal of Political Economy*, 1994, vol. 102, no. 5]

and property rights. Employers, for instance, must decide how much to spend on monitoring workers and how to penalize dysfunctional behavior (Lazear 1986, 1991; Dickens et al. 1989).

Most previous research on these issues has assumed that each individual chooses whether or not to commit one act (the "single-act" framework) (see, e.g., Becker 1968; Landes and Posner 1975; Polinsky and Shavell 1984; Friedman 1981). Realistically, however, truckers can overload a little or a lot, factories choose among many degrees of pollution, and even burglars decide how many houses to raid. In such contexts, of *marginal* deterrence (Stigler 1970), stepping up enforcement against one level of the activity may induce a switch to a more harmful act instead. Friedman and Sjostrom (1991), Mookherjee and Png (1992), Shavell (1991, 1992), and Wilde (1992) study the conditions under which marginal deterrence requires penalties to be graduated. Here, we focus on the *optimal* pattern of marginal deterrence as a function of enforcement costs. Our setting is fairly general: the level of the activity is a continuous variable, and individuals derive heterogeneous benefits. We assume that the monitoring system detects all acts, regardless of their harmfulness, at the same rate. Acts of differing severity may, however, be prosecuted and penalized at different rates.

Our first set of results supposes that prosecution and punishment are costless and only monitoring is costly. Under these assumptions, Friedman (1981) proved that the *average* expected penalty should fall short of average harm.[1] Here, we show that it is optimal to set marginal expected penalties *everywhere* less than marginal harm; that is, society should impose uniformly less than the first-best pattern of deterrence. Suppose that an enforcement policy involves marginal expected penalties equal to or exceeding marginal harms at some point. If society then reduces monitoring, individuals choosing acts in that neighborhood will graduate to more harmful choices; this reduces welfare, but only by a second-order amount. The cut in monitoring, however, generates a first-order welfare gain; hence the original enforcement policy could not be optimal.[2]

Not only should marginal expected penalties be everywhere less than marginal harm, but there should be no enforcement at all against acts below some *threshold*. Generally, there are two ways to deter individuals from causing greater harm. One is to raise the penalties for the more harmful acts, and the other is to reduce the penal-

[1] Strictly, this argument is an extension of a result in his single-act framework (see also Polinsky and Shavell 1984).

[2] See Sec. III for the argument in detail.

ties for or, at the extreme, legalize the less harmful acts. Such legalization reduces the cost of deterring the *greater* harms.[3]

Our next result is that, if prosecution and punishment as well as monitoring are costly, then marginal expected penalties for minor acts should be still lower. In particular, the enforcement threshold should be higher. On the other hand, marginal expected penalties for more serious acts should be stiffer. The idea is to induce individuals to shift away from acts that involve higher prosecution and punishment costs. Since these costs rise with the penalty imposed and more harmful acts require heavier penalties, the optimal enforcement policy encourages individuals to shift from more to less harmful acts.

Accordingly, in general, for minor acts, marginal expected penalties should always fall short of marginal harm; for more harmful acts, the relation between expected penalties and harm depends on the balance between the costs of monitoring and the costs of prosecution/punishment. If the latter costs are sufficiently large, then marginal expected penalties should exceed marginal harm for a range of the most harmful acts.

Our final set of results concerns how the optimal enforcement policy varies with changes in enforcement costs. If monitoring becomes more costly, marginal (and thus also total) expected penalties should be lower at every level. In part, this means less monitoring. It also means lower penalties for less harmful acts and, in particular, a higher enforcement threshold. On the other hand, if prosecution and punishment become more costly, expected penalties ought to be reduced for less harmful acts and *raised* for more harmful ones.

These findings differ sharply from some conclusions based on the single-act framework. Consider the issue of which harmful actions *should* be legalized. According to Posner (1992, p. 224), those that convey a private benefit exceed the external harm. For example, a traveler stranded in a forest should be allowed to steal food from an unattended shack. We contend, however, that Posner's answer is *underinclusive*. It may be optimal to permit some acts, although their marginal benefits fall short of marginal harms, in order to reduce the costs of deterring greater harm.

To analyze the optimal relation between penalties and harms, Polinsky and Shavell (1992) extended the single-act framework in the following way. While *each* individual can commit either one particular act or none at all, different individuals are capable of different acts. Polinsky and Shavell prove that "the optimal fine equals the harm, properly inflated for the chance of not being detected, plus the vari-

[3] This was first shown by Shavell (1992) and Wilde (1992) in discrete contexts.

able enforcement cost of imposing the fine" (p. 133).[4] This rule implies, for instance, that marginal expected penalties should *always* exceed marginal harm, expected penalties should be *independent* of monitoring costs, and if prosecution/punishment costs are higher, expected penalties should be raised for *all* acts.[5]

In light of these differences, we believe that it is important to understand the nature of optimal enforcement in a marginal deterrence framework. Section II presents the general setting. We derive the results, first assuming that prosecution and punishment are costless (Sec. III) and then that they are costly (Sec. IV). Section V concludes the paper.

II. General Setting

Each individual can choose the degree, denoted $a \geq 0$, of some act. Different persons, however, receive different (private) benefits.[6] We represent the heterogeneity by a parameter t, where a type t individual derives benefit $tb(a)$ from act a, so at every level of the act, a higher type derives larger total and marginal benefit.[7] Let the types be distributed according to a positive and continuous density, $g(t)$, on an interval $[0, T]$. The function $b(a)$ relates private benefits to the alternative levels of harm. We assume that $b(\cdot)$ is differentiable and strictly increasing in a. Let $\lim_{a \to \infty} b(a) = \bar{b}$. Since we do not allow infinitely large punishments, we assume that $\bar{b}$ is finite; otherwise it would be impossible to secure any deterrence at all.

The act a imposes external harm $h(a)$, where the function $h(\cdot)$ is differentiable and strictly increasing in a. We should emphasize that the a could instead represent several distinct actions. What is necessary is that if a causes more harm than a', then a must provide more

[4] Strictly, as Polinsky and Shavell note, the optimal fine is the lesser of this formula and the individual's wealth.

[5] In Polinsky and Shavell's setting, the margin of deterrence is *extensive*. Changes in enforcement against some act affect only those individuals who are capable of committing that act. By contrast, we focus on an *intensive* margin of deterrence: each individual is capable of every act, so any change in enforcement potentially affects everyone.

[6] If all persons were identical, it would be optimal to impose the same act on everyone. Then the optimal penalty structure would be a step function: the regulator should legalize all acts up to the optimal level and set maximal penalties for all more harmful acts. Moreover, in equilibrium, no one need actually be punished, so costs of prosecution and punishment would be irrelevant. We avoid this unrealistic scenario through the assumption of a heterogeneous population.

[7] Our main results would not be qualitatively affected if we were to let benefits take a more general structure, as long as a higher type derives uniformly higher total and marginal benefits. This single-crossing condition is used in most self-selection analyses, e.g., Cooper (1984) and Maskin and Riley (1984). See, however, Srinagesh, Bradburd, and Koo (1992).

(private) benefits to all individuals.[8] An example of the setting that we have in mind is long-distance trucking. Each trucker decides on his vehicle load, and different truckers may derive different benefits from the same load. Ceteris paribus, the damage to the road surface rises with the vehicle load.

For simplicity, we assume that all parties are risk-neutral and adopt a utilitarian approach, that is, one that attaches equal weight to private benefits, external harms, and enforcement costs.[9] Accordingly, the "first-best" actions a_t^* (i.e., those that balance each individual's marginal benefit against the corresponding marginal harm) satisfy

$$tb'(a_t^*) = h'(a_t^*). \tag{1}$$

If enforcement were costless and the regulator could distinguish individuals' types, each individual should be compelled to choose his or her respective a_t^*.

To deter overloading, the state highway authority opens roadside weighing stations at random. Enforcement has three aspects: detecting, prosecuting, and punishing offenders. A station will detect minor and major overloaders at a common rate, say μ. The highway authority sets a policy of prosecuting a fraction $p(a)$ of all truckers detected to have taken action a and imposing a corresponding penalty of monetary value $f(a)$. The penalty, which could be pecuniary or nonpecuniary, is subject to an exogenous maximum, w.[10] For simplicity, we assume that there are no errors in enforcement. We also assume, as seems realistic, that the regulator lacks the information to condition enforcement directly on a trucker's type. Accordingly, an enforcement policy consists of a monitoring rate, μ, prosecution rates, $p(a)$, and penalties, $f(a)$, subject to the constraints

$$0 \leq \mu \leq 1, \quad 0 \leq p(a) \leq 1, \tag{2}$$

$$0 \leq f(a) \leq w, \quad f(0) = 0. \tag{3}$$

[8] Friedman and Sjostrom (1991) discuss enforcement of marginal deterrence when this assumption does not hold.

[9] As we explain below, our main results are not sensitive to this assumption. They extend to the case in which arbitrary nonnegative weights are applied to private benefits and social harms. In particular, they apply to contexts in which benefits do not count at all.

[10] Like Becker (1968), Stigler (1970), and others, we implicitly assume that the regulator can set the enforcement rate independently of the penalties. See Malik (1990) and Andreoni (1991) for critiques of this premise. In the case of pecuniary penalties, this maximum might represent the truckers' (identical) wealth. The assumption of identical wealth is not essential. Our analysis actually pertains to all individuals with a particular wealth level. We can perform the analysis separately for each wealth group. The only necessary modification is that a common monitoring rate must apply to individuals with different wealth. But this will not affect our qualitative results.

Given some policy, type t will choose a_t to maximize

$$tb(a) - \mu p(a) f(a) = tb(a) - e(a), \tag{4}$$

where $e(a) \equiv \mu p(a) f(a)$ denotes the expected penalty on act a. We say that the schedule of actions, a_t, is *implemented* by the enforcement policy $\{\mu, p(\cdot), f(\cdot)\}$ if it maximizes (4) for all t.

We now specify the nature of enforcement costs. The costs of monitoring will depend mainly on the number of trucks inspected. For simplicity, we assume that the cost of monitoring at rate μ is $c_M \mu$ per truck, so the total monitoring cost is

$$c_M \mu \int_0^T g(t)\,dt = c_M \mu.$$

By contrast, the costs of prosecution and punishment will depend not only on the number of trucks prosecuted but also on the penalty. Let $c_P(f)$, where $c_P'(\cdot) > 0$, be the cost of prosecuting and imposing a penalty of monetary value f on one individual. Further, let the total cost of prosecution and punishment be simply the sum of the costs for each individual, that is,

$$\mu \int_0^T p(a_t) c_P(f(a_t)) g(t)\,dt,$$

where a_t represents the schedule of actions chosen by the various types.

Having laid out the setting, we are ready to describe the optimal enforcement policy. This policy maximizes

$$W = \int_0^T [tb(a_t) - h(a_t)] g(t)\,dt - \mu c_M - \mu \int_0^T p(a_t) c_P(f(a_t)) g(t)\,dt \tag{5}$$

subject to the constraints that the policy satisfies (2) and (3) and that it implements the schedule a_t.

Analytically, the most convenient approach is to treat the schedule of actions, a_t, as the principal choice variable. The class of feasible schedules is subject to two restrictions. First, higher types cannot be compelled to choose less harmful acts; that is, a_t must be nondecreasing, essentially because higher types derive greater marginal and total benefits than lower types. Consider types s and t ($s < t$). By (4), $tb(a_t) - e(a_t) \geq tb(a_s) - e(a_s)$ and $sb(a_s) - e(a_s) \geq sb(a_t) - e(a_t)$. Adding these two conditions, we obtain $(t - s)[b(a_t) - b(a_s)] \geq 0$, which implies that $a_t \geq a_s$.

The second restriction arises from the limit on penalties. We give a heuristic argument here and leave a rigorous proof to the Appendix. Define the indirect (maximized) utility function, $V(t) \equiv tb(a_t) - e(a_t)$.

Then by the envelope theorem, if the schedule a_t maximizes utility for each type,

$$V'(t) = b(a_t). \tag{6}$$

Assuming that the schedule a_t is continuous,[11] we can integrate (6) and substitute for V to obtain

$$tb(a_t) - e(a_t) = \int_0^t b(a_\tau)d\tau.$$

Rearranging, we have the expected penalties necessary to implement the schedule a_t,

$$e(a_t) = tb(a_t) - \int_0^t b(a_\tau)d\tau. \tag{7}$$

Condition (7) specifies expected penalties for all acts actually chosen by some type; that is, $a \in [a_0, a_T]$, where a_0 and a_T are the acts selected by the lowest and highest types, respectively. To implement the schedule a_t, the enforcement policy must also deter all individuals from acts more severe than a_T. Since benefits increase with type t, a sufficient condition is that the policy deter the highest type T from all $a > a_T$, that is, $Tb(a) - e(a) \leq Tb(a_T) - e(a_T)$, all $a > a_T$. Rearranging, we have $e(a) \geq T[b(a) - b(a_T)] + e(a_T)$, which sets a lower bound to the expected penalties on acts $a > a_T$. Without loss of generality, the expected penalty can be set equal to this lower bound. Then, substituting for $e(a_T)$ from (7) above, we obtain

$$e(a) = Tb(a) - \int_0^T b(a_\tau)d\tau \tag{8}$$

for all $a > a_T$.

Now the expected penalty on any act cannot exceed the maximum possible punishment w. Let $a \rightarrow \infty$; hence it follows that a schedule a_t can be implemented only if it satisfies

$$w \geq T\bar{b} - \int_0^T b(a_\tau)d\tau, \tag{9}$$

which is the second restriction on the set of feasible schedules a_t. Notice that, by (9), if society seeks more deterrence (i.e., to impose lower a_τ), then $b(a_\tau)$ will be lower; thus the right-hand side of (9) will be higher. Accordingly, the maximum possible penalty, w, sets a limit to how much deterrence is feasible. The essential reason is that, by (7), if society wishes to reduce some a_τ, it must raise expected penalties for

[11] We show below that the optimal schedule will be continuous.

all more harmful acts, a_t, for $t > \tau$. Expected penalties, however, cannot be raised beyond w. This is the key to understanding the results of the following section.

We have shown, heuristically, that a pattern of behavior, a_t, can be implemented only if it is nondecreasing and meets (9). In the Appendix, we prove formally that these conditions are necessary and also sufficient. Accordingly, in seeking the optimal enforcement policy, we can limit attention to policies inducing choices that satisfy these two conditions.[12]

Lemma. A schedule of choices, a_t, can be implemented by some enforcement policy if and only if a_t is nondecreasing and it satisfies equation (9):

$$w \geq T\bar{b} - \int_0^T b(a_\tau)\,d\tau.$$

The requisite expected penalties are

$$e(a) = t(a)b(a) - \int_0^{t(a)} b(a_\tau)\,d\tau, \tag{10}$$

where $t(a)$ denotes the highest (supremum) type τ selecting an $a_\tau \leq a$.

By the lemma, the expected penalties necessary to enforce an implementable pattern of behavior a_t are

$$e(a) = \mu p(a) f(a) = t(a)b(a) - \int_0^{t(a)} b(a_\tau)\,d\tau.$$

From these penalties we can derive the corresponding monitoring rate. By (2) and (3), respectively, $1 \geq p(a)$ and $w \geq f(a)$; hence the monitoring rate must satisfy

$$\mu w \geq t(a)b(a) - \int_0^{t(a)} b(a_\tau)\,d\tau \quad \text{for all } a.$$

Since this lower bound is increasing in a, we obtain an equivalent condition by letting $a \rightarrow \infty$:

$$\mu \geq \frac{1}{w}\left[T\bar{b} - \int_0^T b(a_\tau)\,d\tau\right]. \tag{11}$$

From equation (5), for welfare from an implementable schedule a_t to be maximized, monitoring must be minimized subject to (2) and the

[12] Strictly, the lemma defines expected penalties only for acts actually chosen, i.e., $a \in [a_0, a_T]$. Provided that expected penalties for $a > a_T$ are severe enough that no type deviates to such a, these penalties will never be applied and so do not affect welfare; likewise, if expected penalties for $a < a_0$ are sufficiently low. Accordingly, subject to this proviso, for simplicity, we can limit attention to penalties on $a \in [a_0, a_T]$.

schedule being implemented. Consider setting

$$\mu = \frac{1}{w}\left[T\bar{b} - \int_0^T b(a_\tau)\,d\tau\right].^{13} \tag{12}$$

Since $b(\cdot)$ is increasing,

$$\frac{T\bar{b} - \int_0^T b(a_\tau)\,d\tau}{w} > 0.$$

Further, by (9), since a_t can be implemented,

$$\frac{T\bar{b} - \int_0^T b(a_\tau)\,d\tau}{w} \leq 1.$$

Thus (12) meets (2) and hence must be the optimal monitoring rate.

III. Costless Prosecution and Punishment

As a first step toward characterizing optimal enforcement policies, we assume in this section that prosecution and punishment are costless ($c_P = 0$). In this case, it is optimal to prosecute every person detected by the monitoring system, that is, set $p(a) = 1$ for all a. If we substitute $c_P = 0$ and (12) in (5), welfare simplifies to

$$W = \int_0^T [tb(a_t) - h(a_t)]g(t)\,dt - \frac{c_M}{w}\left[T\bar{b} - \int_0^T b(a_\tau)\,d\tau\right].$$

So, by the lemma, the regulator seeks a schedule of actions a_t and an enforcement policy to maximize

$$W = \int_0^T \left\{\left[t + \frac{c_M}{wg(t)}\right]b(a_t) - h(a_t)\right\}g(t)\,dt - \frac{T\bar{b}c_M}{w} \tag{13}$$

subject to a_t being nondecreasing and (9).

[13] We reiterate that it is optimal for constraint (11) to bind even though prosecution and punishment are costly. Implementation of some a_t defines a schedule of expected penalties, $e(a) \equiv \mu p(a)f(a)$. If the regulator were to increase monitoring, she could then proportionately reduce prosecution rates without affecting expected penalties. Prosecution costs,

$$\mu \int_0^T p(a_t)c_P(f(a_t))g(t)\,dt,$$

however, would remain unchanged. So society would be worse off by the increase in monitoring costs. Accordingly, monitoring should be kept to the minimum.

Let ν/w be the Lagrange multiplier on (9). Then consider maximizing pointwise

$$\left[t + \frac{c_M + \nu}{wg(t)}\right] b(a) - h(a). \tag{14}$$

Suppose that the density $g(t)$ is such that $t + [(c_M + \nu)/wg(t)]$ is everywhere nondecreasing in t, as will be true if t is uniformly distributed. Then the pointwise solution defines a continuous, nondecreasing schedule a_t and, hence, solves the regulator's problem.[14] The schedule a_t defined by (14) is continuous because the density function $g(t)$ is continuous. Having found the optimal schedule a_t, we must derive the enforcement policy that implements it at minimum cost. The monitoring rate is given by (12). Since $p(a) = 1$, the requisite expected penalties $e(a) = \mu f(a)$; hence, by (12) and (10), the actual penalties must satisfy

$$f(a) = w\left[\frac{t(a) b(a) - \int_0^{t(a)} b(a_\tau)\, d\tau}{T\bar{b} - \int_0^T b(a_\tau)\, d\tau}\right]. \tag{15}$$

We next describe several key qualitative properties of the optimal enforcement policy. By (14), if the optimal act $a_t > 0$, then it must satisfy the first-order condition

$$\left[t + \frac{c_M + \nu}{wg(t)}\right] b'(a_t) = h'(a_t). \tag{16}$$

Since $c_M > 0$ and $\nu \geq 0$, (16) and (1) imply that $a_t \geq a_t^*$;[15] that is, the regulator should enforce *less* than the first-best degree of deterrence on *all* types. To enforce this, marginal expected penalties should be less than the corresponding marginal harms at all levels of the activity actually chosen. To see this, note that, by (4) with $p(a) = 1$, for the various types to voluntarily choose the schedule a_t, it must satisfy the first-order condition

$$tb'(a_t) = \mu f'(a_t). \tag{17}$$

[14] Below, we shall argue that our results are not qualitatively changed if the density does not meet this condition for all t.

[15] The inequality is strict if the first-best act $a_t^* > 0$.

Substituting in (16) above, we get

$$\mu f'(a_t) = \left[1 + \frac{c_M + \nu}{wtg(t)}\right]^{-1} h'(a_t), \tag{18}$$

which implies that $\mu f'(a_t) < h'(a_t)$, as claimed.[16]

The single-act analyses of Friedman (1981) and Polinsky and Shavell (1984) suggest why it is not optimal to match or exceed the first-best pattern of deterrence everywhere. Suppose otherwise that all choices are first-best. Then society should reduce monitoring. Individuals will shift toward more harmful acts, but since their original choices were first-best, their incremental benefits will almost balance the corresponding incremental harms. The reduction in monitoring, however, will cut monitoring costs by a first-order amount, unequivocally raising welfare. The same argument applies a fortiori when the original choices are less harmful than first-best.

In the marginal deterrence setting, we prove that it is not optimal to match or exceed the first-best degree of deterrence for *any* type. The key is to realize that the limit on punishments, w, constrains how much deterrence (costless) penalties alone can provide. Without such a constraint, any desired pattern of deterrence could be achieved at minimal cost by combining arbitrarily low monitoring with sufficiently steep penalties.[17]

Given the limit on punishments, w, suppose that an enforcement policy involves marginal expected penalties equal to marginal harms at some point. Consider first reducing penalties at that and all higher points. Individuals choosing acts in that neigborhood will graduate to more harmful choices, reducing welfare, but only by a second-order amount. For others, whose original choices were further away, the variation will be inframarginal; hence they will not change their actions. The variation reduces the difference between the penalties for the most and least harmful acts to less than w. Society should then use this freedom to raise penalties everywhere and reduce the monitoring rate, while preserving *marginal* expected penalties. The cut in monitoring generates a first-order welfare gain; thus the original enforcement policy could not be optimal.

[16] This argument applies to all acts actually chosen by some type. The same also holds for all less harmful acts, $a < a_0$, where a_0 represents the choice of type 0. For such $a < a_0$, expected penalties should be set to zero, so marginal expected penalties will be zero and will be clearly less than marginal harm. Regarding $a > a_T$, penalties converge to w as $a \rightarrow \infty$; hence marginal expected penalties must eventually vanish. By contrast, marginal harms will typically be positive, so for $a > a_T$, marginal expected penalties will eventually fall below marginal harms.

[17] Friedman (1981) and Polinsky and Shavell (1984) also assume that punishments are limited.

The same argument applies, a fortiori, if marginal expected penalties originally exceed the corresponding marginal harms at some point: in this case, allowing individuals in that neighborhood to cause more harm will itself generate a first-order welfare gain.

Our results are not substantially affected even if $t + [(c_M + v)/wg(t)]$ falls with t over one or more intervals $[t', t'']$. In this case, (14) characterizes the optimal acts for types t' and t'', and all intermediate types will be pooled at the common act $a = a_{t'} = a_{t''}$. There will be a kink in the expected penalty function at a to effect this pooling of types. The techniques of Myerson (1981) and Maskin and Riley (1984), however, show that such pooling does not affect the qualitative properties of the optimal enforcement policy. For instance, by (14), it will be optimal to allow type t'' to cause more than its first-best degree of harm, that is, $a_{t''} > a^*_{t''}$. Since $a_{t''} = a$, all $t \in [t', t'']$, and, by (1), $a^*_{t''} > a^*_t$, all $t < t''$, it follows that $a_t > a^*_t$, all $t \in [t', t'']$; that is, it will be optimal to allow all the other types within the interval $[t', t'']$ to cause more than their first-best degree of harm as well. Accordingly, from now on, we ignore the constraint that a_t be nondecreasing.

Our next result is that the regulator should legalize all acts below some threshold, a_0. By (14), the optimal action for type t arbitrarily close to zero maximizes

$$\left[\frac{c_M + v}{wg(0)}\right] b(a) - h(a). \tag{19}$$

If monitoring is sufficiently costly in the sense that

$$\left[\frac{c_M + v}{wg(0)}\right] b'(0) > h'(0), \tag{20}$$

then this type should choose some $a_t > 0$. To induce such a choice, the expected penalties for all less harmful acts should be zero. By legalizing acts below some threshold, the regulator causes more harmful acts (above the threshold) to become relatively less attractive and, hence, can deter them at a lower cost.

We can also say how optimal enforcement depends on the maximum possible punishment, w. An increase in w would reduce v and so lower the left-hand side of (16), which means that society should *step up* deterrence. Intuitively, the higher w increases the scope for deterrence through penalties; hence it is optimal to adjust accordingly. In the Appendix, we prove that a fall in monitoring costs, c_M, would also reduce the left-hand side of (16). Intuitively, when monitoring (hence deterrence) becomes less costly, society should move closer to the first-best pattern of deterrence.

PROPOSITION 1. Under the enforcement policy that is optimal when prosecution and punishment are costless,

a) provided that monitoring is sufficiently costly, some range of less harmful acts should be legalized;

b) the marginal expected penalty should be strictly less than the corresponding marginal social harm for any act actually chosen; and

c) if the maximum possible punishment, w, is lower or the cost of monitoring, c_M, is higher, the regulator should (i) reduce the monitoring rate, (ii) raise the enforcement threshold, and (iii) reduce the expected penalty on all more harmful acts chosen, in both absolute and marginal terms.

To illustrate the foregoing results, consider the following example. Let $b(a) = \bar{b} - e^{-\beta a}$, $h(a) = ha$, with $\beta, h > 0$, and t be uniformly distributed on [0, 1]. Suppose that monitoring is sufficiently costly that, under the optimal pattern of choices, a_t, there will be an enforcement threshold, and, further, constraint (9) will not bind, in which case, $\nu = 0$.[18] If we substitute in (16), the optimal actions are

$$a_t = \frac{1}{\beta} \ln\left[\frac{\beta}{h}\left(t + \frac{c_M}{w}\right)\right]. \tag{21}$$

Hence, all levels of the activity up to

$$a_0 = \frac{1}{\beta} \ln\left(\frac{\beta c_M}{hw}\right) \tag{22}$$

should be legalized. If we substitute (21) in (12) and (15), the optimal monitoring rate and penalties are

$$\mu = \frac{h}{\beta w} \ln\left(1 + \frac{w}{c_M}\right) \tag{23}$$

and

$$f(a) = \begin{cases} 0 & \text{if } a \leq a_0, \\ \dfrac{1}{\mu}\left(ha + \dfrac{c_M}{w} e^{-\beta a} - \dfrac{h}{\beta} + \dfrac{h}{\beta} \ln \dfrac{hw}{\beta c_M}\right) & \text{if } a_0 < a < a_1, \\ w - \dfrac{1}{\mu} e^{-\beta a} & \text{if } a \geq a_1. \end{cases} \tag{24}$$

[18] A sufficient condition is that $c_M/w > \max\{h/\beta, [\exp(w\beta/h) - 1]^{-1}\}$.

Now let $\beta = 5$, $h = 1$, and $w = 20$, and focus on the effect of increases in the cost of monitoring, c_M, on the optimal enforcement policy. Figure 1*a* graphs optimal expected penalties when $c_M = 0$, 5, and 10, respectively.[19] The expected penalty function when $c_M = 0$, the first-best case, coincides with the external harm. When $c_M > 0$, the optimal expected penalty is zero up to the enforcement threshold a_0 and then begins to rise.[20] Note that the higher the cost of monitoring, the higher the enforcement threshold and the lower and flatter the expected relative function should be. By contrast, the single-act analyses of Landes and Posner (1975) and Polinsky and Shavell (1992) prescribe marginal expected penalties equal to marginal harm plus enforcement costs; thus they are clearly higher than marginal harm. This is one of the key results that distinguish single-act from marginal deterrence settings.

In figure 1*b*, we graph the corresponding choices of action. The higher the monitoring cost, the less deterrence society should try to impose. This is represented by the schedule a_t shifting uniformly upward. Notice that as c_M rises, the optimal action a_t increases relatively faster for lower types.[21] Next, figure 1*c* shows that, as the cost of monitoring rises, it is optimal to uniformly *reduce* penalties. By contrast, in the framework of Polinsky and Shavell (1992), to the extent that an increase in the cost of monitoring makes it desirable to reduce monitoring, *all* penalties should be *raised.*

The example highlights yet another key distinction. If marginal expected penalties equal marginal harm plus enforcement costs, then they will be independent of the pattern of private benefits. In the marginal deterrence framework, however, the optimal enforcement policy clearly does depend on private benefits (see [12] and [15] above). Figure 2 plots optimal expected penalty functions for $\beta = 5$, 10, and 15 with $c_M = 5$. As may be seen, the effects of raising marginal benefits are quite complex. On the one hand, as β increases, each type derives more benefit from the activity. Since benefits count

[19] The expected penalty functions in the figure have a second parameter, namely, the cost of prosecution and punishment, which, for the moment, is set equal to zero.

[20] The penalties on acts more harmful than those chosen by the highest type, i.e., $a > a_1$, are not unique. For instance, the optimal schedule of actions can also be implemented by penalties that jump to w for $a > a_1$ or any increasing function that lies between w and the function graphed in fig. 1*a*. In the figure, we have used the lowest penalties on $a > a_1$ sufficient to implement a_t. This also ensures that the schedule $f(a)$ will be continuous.

[21] By (17), individuals' choices are guided by $tb'(a_t) = \mu f'(a_t)$. As monitoring costs rise, the cost of enforcing a given pattern of behavior a_t rises. To economize, it is optimal to compress the range of acts actually chosen, thereby reducing the lower types' incentive to switch upward to more harmful acts.

in social welfare, this is an argument for allowing more harm, that is, weaker enforcement. On the other hand, ceteris paribus, higher marginal benefits induce all persons to choose more harmful acts. So unless enforcement is strengthened, all types will gravitate toward causing more harm.

IV. Costly Prosecution and Punishment

We turn next to study optimal enforcement when prosecution and punishment as well as monitoring are costly. Recall that the regulator seeks a schedule of choices, a_t, and an enforcement policy to maximize (5) subject to a_t being nondecreasing and meeting (9). Before characterizing the optimal choices, we solve for the optimal enforcement policy as a function of an implementable schedule of choices. From (12), we have the monitoring rate. It remains to derive the prosecution rates and the penalties.

Referring to (5), to maximize welfare from an implementable schedule a_t, we must, for each type t, minimize prosecution and punishment costs

$$\mu p(a_t) c_P(f(a_t)) g(t) \tag{25}$$

subject to (2) and (3) and provided that expected penalties are sufficient to implement the schedule. By (10), the latter condition is

$$\mu p(a_t) f(a_t) \geq tb(a_t) - \int_0^t b(a_\tau) d\tau.$$

To minimize prosecution and punishment costs, prosecution rates and penalties should be set so that this constraint binds; hence

$$\mu p(a_t) f(a_t) = \left[tb(a_t) - \int_0^t b(a_\tau) d\tau \right]. \tag{26}$$

Substituting in (25), the regulator seeks to minimize

$$\mu p(a_t) c_P(f(a_t)) g(t) = \frac{c_P(f(a_t))}{f(a_t)} \left[tb(a_t) - \int_0^t b(a_\tau) d\tau \right] g(t) \tag{27}$$

subject to (3), for all t.

Let $\hat{f}$ minimize $c_P(f)/f$ subject to (3). This "efficiency penalty" solves (27) for all t.[22] By (26), to obtain the requisite expected penalty, the

[22] The steeper the marginal prosecution and punishment cost, $c_P'(\cdot)$, the smaller the efficiency penalty $\hat{f}$.

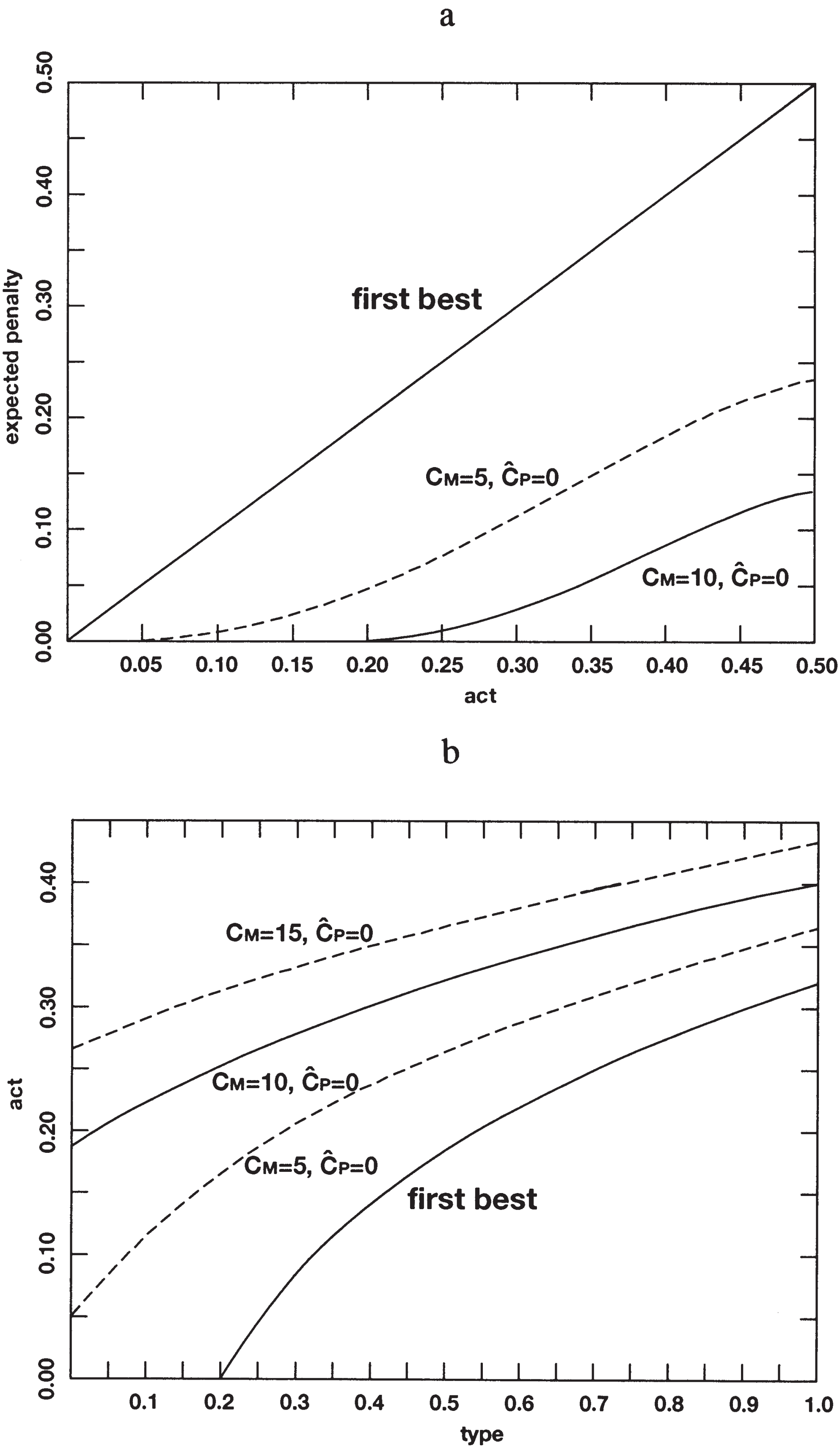

FIG. 1.—*a*, Expected penalties as monitoring costs change. *b*, Individual choices as monitoring costs change. *c*, Penalties as monitoring costs change.

c

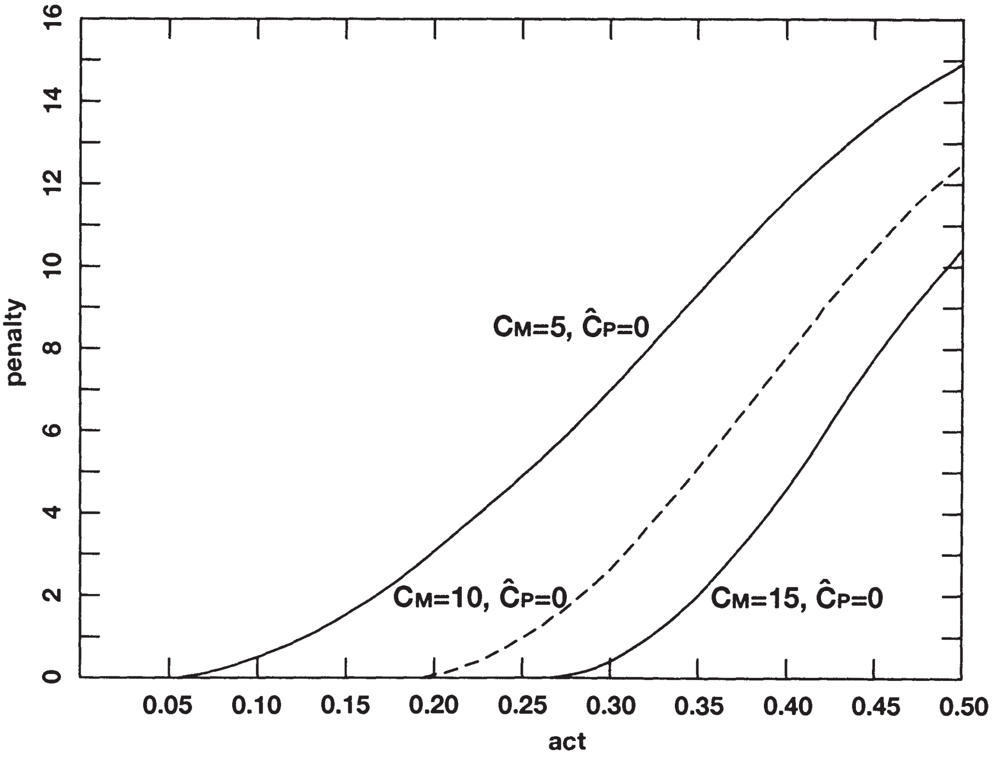

FIG. 1.—*Continued*

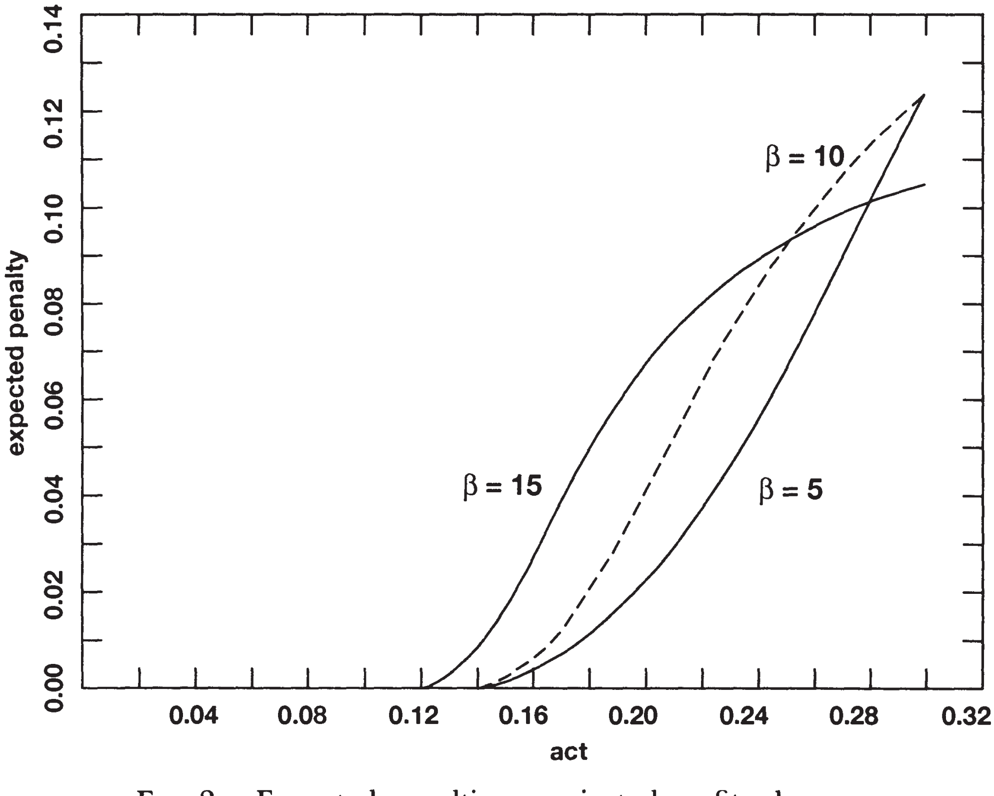

FIG. 2.—Expected penalties as private benefits change

regulator should vary the prosecution rate,[23]

$$p(a_t) = \frac{tb(a_t) - \int_0^t b(a_\tau)\,d\tau}{\mu\hat{f}}. \tag{28}$$

Let $\hat{c}_P$ represent the cost of prosecution and imposing the optimal punishment on one individual, that is, $\hat{c}_P \equiv c_P(\hat{f})$. If we substitute (12), (27), and the optimal penalty in (5), welfare is

$$\begin{aligned} W = \int_0^T [tb(a_t) - h(a_t)]g(t)\,dt - \frac{c_M}{w}\left[T\bar{b} - \int_0^T b(a_\tau)\,d\tau\right] \\ - \frac{\hat{c}_P}{\hat{f}}\int_0^T \left[tb(a_t) - \int_0^t b(a_\tau)\,d\tau\right]g(t)\,dt. \end{aligned} \tag{29}$$

Let G denote the distribution function of t. Then, if we integrate by parts and rearrange terms, welfare is

$$\begin{aligned} W = \int_0^T \left(\left\{t + \frac{c_M}{wg(t)} - \frac{\hat{c}_P}{\hat{f}}\left[t - \frac{1 - G(t)}{g(t)}\right]\right\}b(a_t) - h(a_t)\right) \\ \times g(t)\,dt - \frac{T\bar{b}c_M}{w}. \end{aligned} \tag{30}$$

The regulator seeks a schedule of choices, a_t, to maximize (30) subject to the schedule being nondecreasing and meeting (9).

We can now characterize the optimal pattern of behavior. As before, let ν/w be the Lagrange multiplier on (9). Then the optimal pattern pointwise maximizes

$$\left\{t + \frac{c_M + \nu}{wg(t)} - \frac{\hat{c}_P}{\hat{f}}\left[t - \frac{1 - G(t)}{g(t)}\right]\right\}b(a) - h(a). \tag{31}$$

[23] If $\hat{f} = w$, then, by (12),

$$\mu\hat{f} = T\bar{b} - \int_0^T b(a_\tau)\,d\tau \geq tb(a_t) - \int_0^T b(a_\tau)\,d\tau,$$

so (28) gives $p(a_t) \leq 1$. If $\hat{f} < w$, for very high a, we may find that

$$tb(a_t) - \int_0^T b(a_\tau)\,d\tau > \mu\hat{f}.$$

In this case, the regulator should set $p(a) = 1$ and graduate the penalty, $f(a)$, from $\hat{f}$ up to the maximum w so as to satisfy (26). This does not affect our key conclusions, which we derive in terms of expected penalties.

If the optimal $a_t > 0$, it will meet the first-order condition

$$\left\{t + \frac{c_M + \nu}{wg(t)} - \frac{\hat{c}_P}{\hat{f}}\left[t - \frac{1 - G(t)}{g(t)}\right]\right\} b'(a) = h'(a). \tag{32}$$

The multiplicand

$$tg(t) - [1 - G(t)] \tag{33}$$

in (32) represents two conflicting effects of costly prosecution and punishment. Suppose that society were to raise a_t, allowing type t to cause more harm. Then, by (28), the regulator must enhance the prosecution rate on a_t by an amount proportional to t. This serves to deter types $\tau < t$ from switching up to the new, higher a_t. The higher prosecution rate applies to all type t individuals. Accordingly, the additional prosecution and punishment costs carry a weight $tg(t)$. On the other hand, if a_t is higher, it will become relatively more attractive to types $\tau > t$. These types will become less inclined toward their original choices, that is, acts more harmful than a_t. By (28), the regulator can uniformly reduce prosecution rates on such $a > a_t$ without affecting marginal deterrence. This reduction in prosecution and punishment costs carries a weight $1 - G(t)$, which is the second term in (33).

Suppose that the distribution of types meets the following three regularity conditions: $g(0)$ is finite, $g(T)$ is positive, and the inverse hazard rate, $[1 - G(t)]/g(t)$, is decreasing in t.[24] The latter condition implies that there will exist some critical value $\hat{t}$ below which $t - \{[1 - G(t)]/g(t)\} < 0$ and above which $t - \{[1 - G(t)]/g(t)\} > 0$. When we compare (32) with (1), the optimal degree of deterrence is unequivocally less than first-best for types $t < \hat{t}$.[25] For these types, the saving on enforcement against acts $a > a_t$ outweighs the increased prosecution and punishment required on the act a_t itself. Indeed, prosecution and punishment costs provide a further reason for an

[24] The uniform distribution satisfies all three conditions.

[25] We emphasize that, even if monitoring costs $c_M = 0$, it is optimal to set marginal expected penalties less than marginal harm for types $t < \hat{t}$. By contrast, the single-act models of Landes and Posner (1975) and Polinsky and Shavell (1992) prescribe penalties that exceed harm by prosecution and punishment costs. Friedman (1981) shows that costly prosecution may imply an optimal expected penalty above or below the social harm, depending on the supply elasticity of the acts. As we show next, the marginal deterrence framework yields a less ambiguous answer: the optimal marginal expected penalty lies below marginal harm only for the less harmful acts, essentially because penalties (hence prosecution and punishment costs) rise with the harmfulness of the act.

enforcement threshold.[26] The threshold act should be set to maximize

$$\left[\frac{c_M + \nu}{wg(0)} + \frac{\hat{c}_P}{\hat{f}g(0)}\right] b(a) - h(a). \tag{34}$$

By contrast, for types $t > \hat{t}$, the balance tips the other way: the increased prosecution and punishment cost required on the act a_t outweighs the saving on enforcement against acts $a > a_t$. The essential reason is that the latter saving depends on the weight $1 - G(t)$ of types $\tau > t$. Obviously, the closer t is to T, the smaller $1 - G(t)$ will be. Accordingly, for types $t > \hat{t}$, the optimal degree of deterrence will *exceed* that when prosecution and punishment are costless.

We now summarize the key results of this section.

PROPOSITION 2. Under the enforcement policy that is optimal when prosecution and punishment are costly,

a) provided that prosecution and punishment are sufficiently costly, some range of less harmful acts should be legalized;
b) provided that the inverse hazard rate of the distribution, $G(t)$, is decreasing in t, there exists a critical type $\hat{t}$ such that the marginal expected penalty should be strictly less than the corresponding marginal social harm for all acts $a < a_{\hat{t}}$; and
c) if the cost of prosecution and punishment, $\hat{c}_P$, is higher, the regulator should (i) raise the enforcement threshold and (ii) provided that the inverse hazard rate is decreasing in t, reduce marginal expected penalties on all acts $a < a_{\hat{t}}$ and raise marginal expected penalties on all more serious acts.[27]

To illustrate the foregoing results, we augment the example of the previous section to include the prosecution and punishment cost, $\hat{c}_P$. For simplicity, we assume that $\hat{c}_P$ is scaled such that the efficiency penalty $\hat{f} = w$. Further, as before, we assume that there will be an enforcement threshold and that (9) does not bind, so that $\nu = 0$.[28]

[26] The existence of a threshold to enforcement is a very robust result. Thresholds seem very common in practice: "If we take away the licence of every incompetent lawyer in New York City, we wouldn't need to recycle the *New York Times*" (Adam Schiff, New York City District Attorney, in *Law and Order,* NBC [January 13, 1993]).

[27] For brevity, we omit the proof of this proposition.

[28] Sufficient conditions are that

$$c_M + \hat{c}_P > \max\left\{\frac{wh}{\beta}, w\left[\exp\left(\frac{w\beta}{h} - 1\right)\right]^{-1}\right\}$$

and that $\hat{c}_P < \frac{1}{2}[w - (h/\beta)]$.

In this case, the optimal schedule of actions is

$$a_t(c_M, \hat{c}_P) = \frac{1}{\beta}\ln\left\{\frac{\beta}{h}\left[t + \frac{c_M}{w} + \frac{\hat{c}_P}{w}(1 - 2t)\right]\right\}. \tag{35}$$

Hence, the enforcement threshold

$$a_0 = \frac{1}{\beta}\ln\left[\frac{\beta(c_M + \hat{c}_P)}{hw}\right],$$

which is increasing in both monitoring and prosecution/punishment costs. By contrast, the choice of the highest type

$$a_1 = \frac{1}{\beta}\ln\left[\frac{\beta}{h}\left(1 + \frac{c_M - \hat{c}_P}{w}\right)\right]$$

depends on the difference between monitoring and prosecution/punishment costs. So whether the highest type's choice exceeds or falls below first-best depends simply on whether monitoring is more or less costly than prosecution and punishment.

For purposes of the following diagrams, we continue to maintain, as in the previous section, $\beta = 5$, $h = 1$, and $w = 20$. Each graph has two arguments, c_M and $\hat{c}_P$. To focus on prosecution and punishment costs, we assume that $c_M = 0$. Figure 3*a* plots optimal expected penalty functions for $c_M = 0$ and $\hat{c}_P = 5, 8$. The higher the prosecution and punishment cost, the uniformly lower the optimal expected penalty. Further, within the range of acts actually chosen by some type, marginal expected penalties are less than marginal harm for less harmful acts and then rise to exceed marginal harms for more harmful ones.

Figure 3*b* presents the corresponding schedules of choices. The most striking aspect of these graphs is that optimal choices neatly divide at $t = 1/2$. Lower types are deterred less than first-best, whereas the opposite holds for higher types. The higher the cost of prosecution and punishment, the narrower the range between the choices of the lowest and highest types. By compressing this range, the regulator reduces the lower types' temptation to switch up to more harmful acts and, hence, can economize on prosecution and punishment. Figure 3*c* presents the corresponding prosecution rates. The higher prosecution and punishment costs, the uniformly lower optimal prosecution rates.

Finally, we illustrate optimal enforcement policies when monitoring as well as prosecution and punishment is costly. Figure 4 presents optimal policies for $(c_M, \hat{c}_P) = (6, 2)$ and $(2, 6)$. As the cost of prosecution/punishment rises relative to the cost of monitoring, optimal ex-

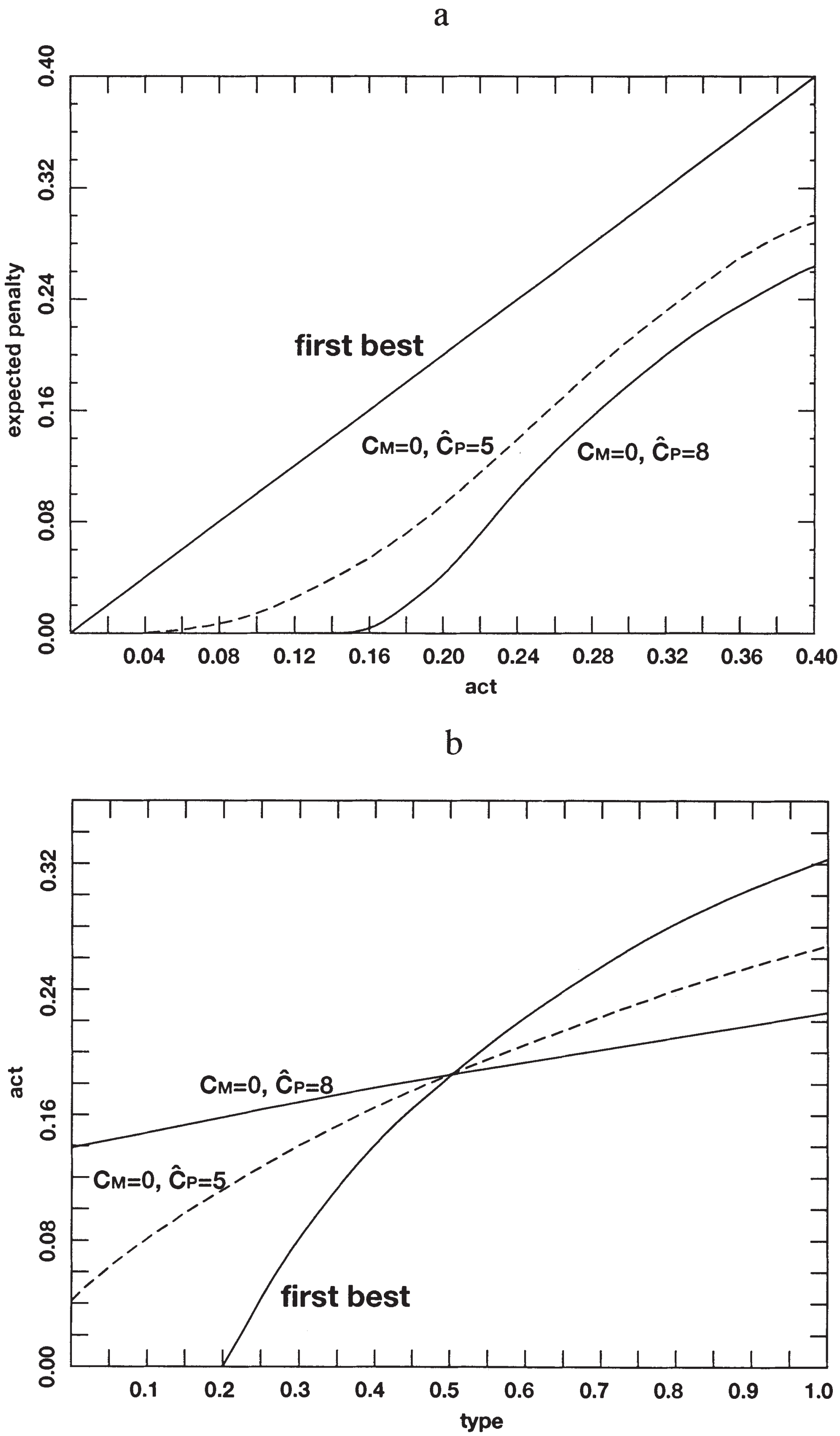

FIG. 3.—*a*, Expected penalties as prosecution/punishment costs change. *b*, Individual choices as prosecution/punishment costs change. *c*, Prosecution/punishment rates as prosecution/punishment costs change.

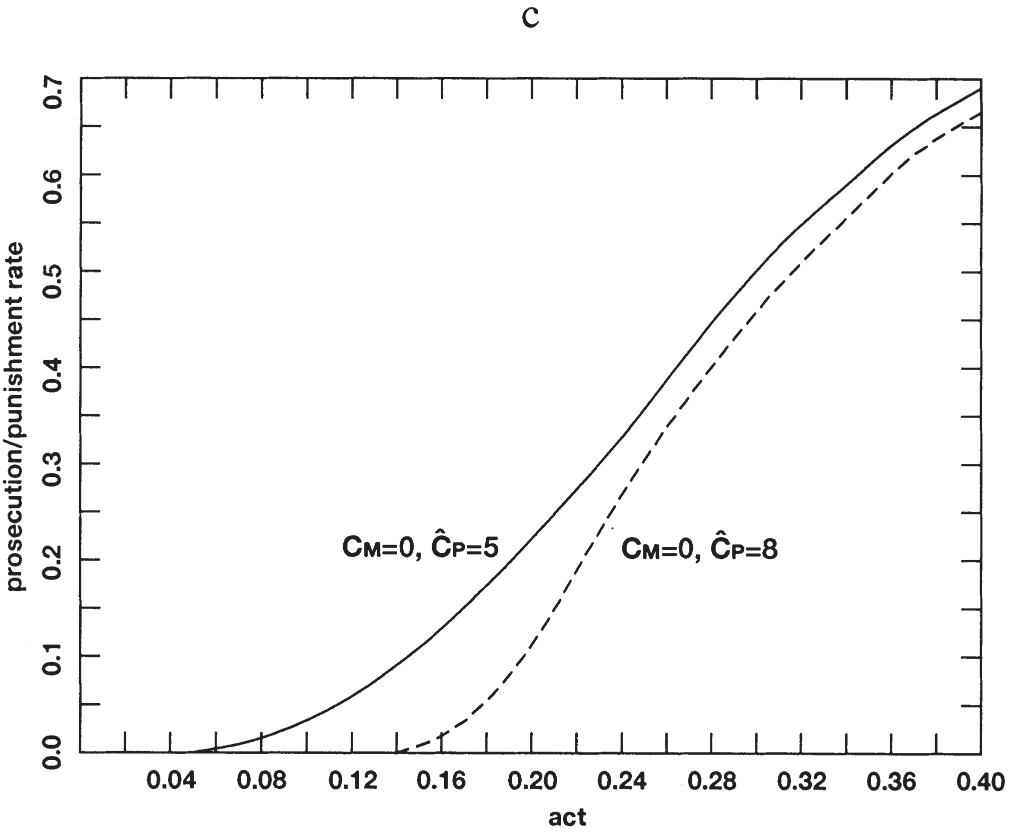

FIG. 3.—*Continued*

pected penalties shift upward, as do marginal expected penalties (fig. 4*a*). Further, it is optimal to compel all types to cause less harm, and, in particular, a range of high types should be deterred beyond first-best (fig. 4*b*). Finally, somewhat surprisingly, prosecution rates should be *higher,* in both absolute and marginal terms (fig. 4*c*). The increased prosecution serves to compress the range between the choices of the lowest and highest types.

V. Concluding Remarks

While we took a utilitarian approach in this paper, we should emphasize that similar results apply when society attaches different weights to private benefits and external harms and, in particular, to acts so reprehensible that society attaches zero weight to private benefits. All that need be done is to adjust the welfare function in (5) to reflect the disparate weights. Then the conditions corresponding to (14) or (32) would show that the lower society's weight on the private benefits from some act, the more it should be deterred and, consequently, the *steeper* should be the schedule of expected penalties.

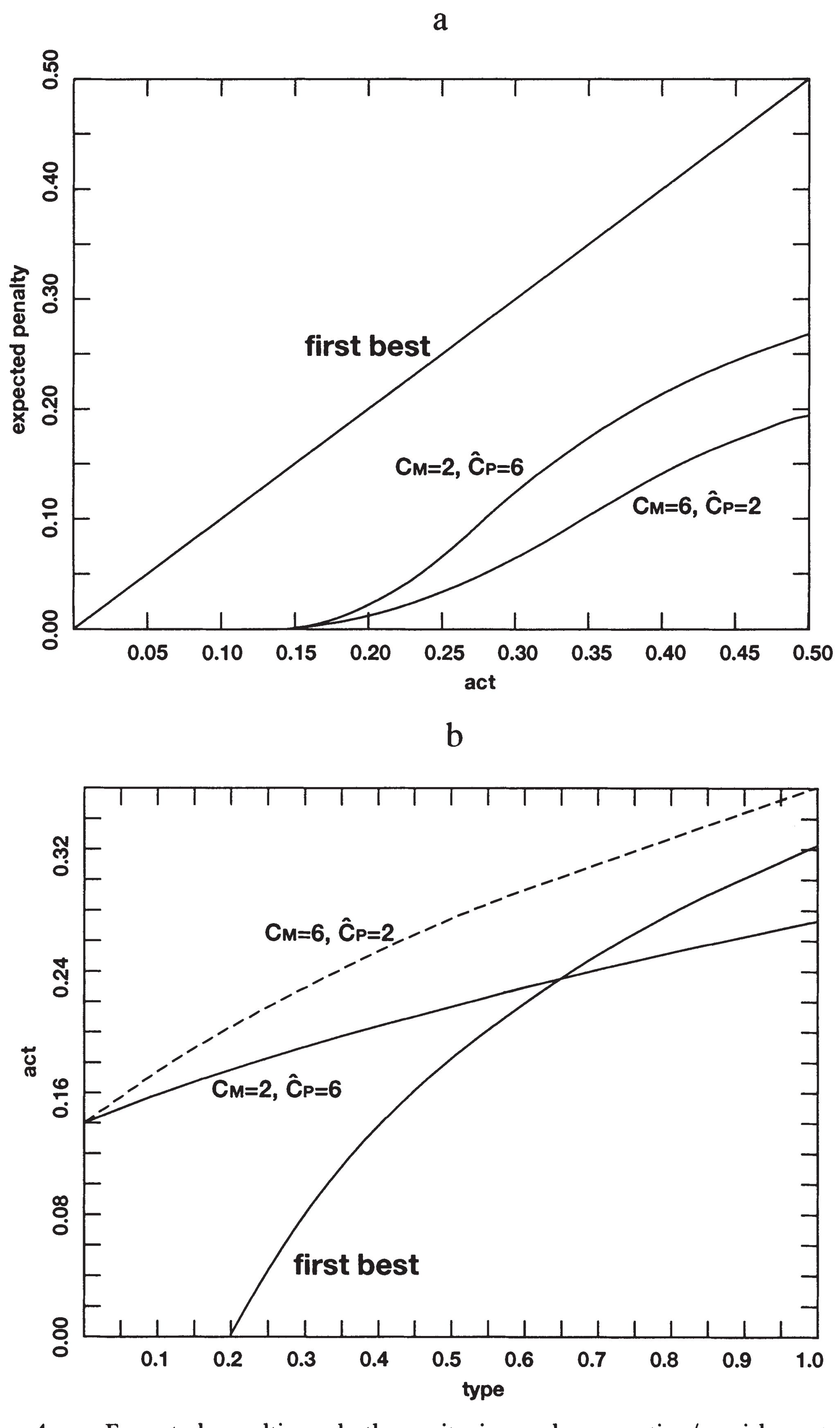

FIG. 4.—*a*, Expected penalties as both monitoring and prosecution/punishment costs change. *b*, Individual choices as both monitoring and prosecution/punishment costs change. *c*, Prosecution/punishment rates as both monitoring and prosecution/punishment costs change.

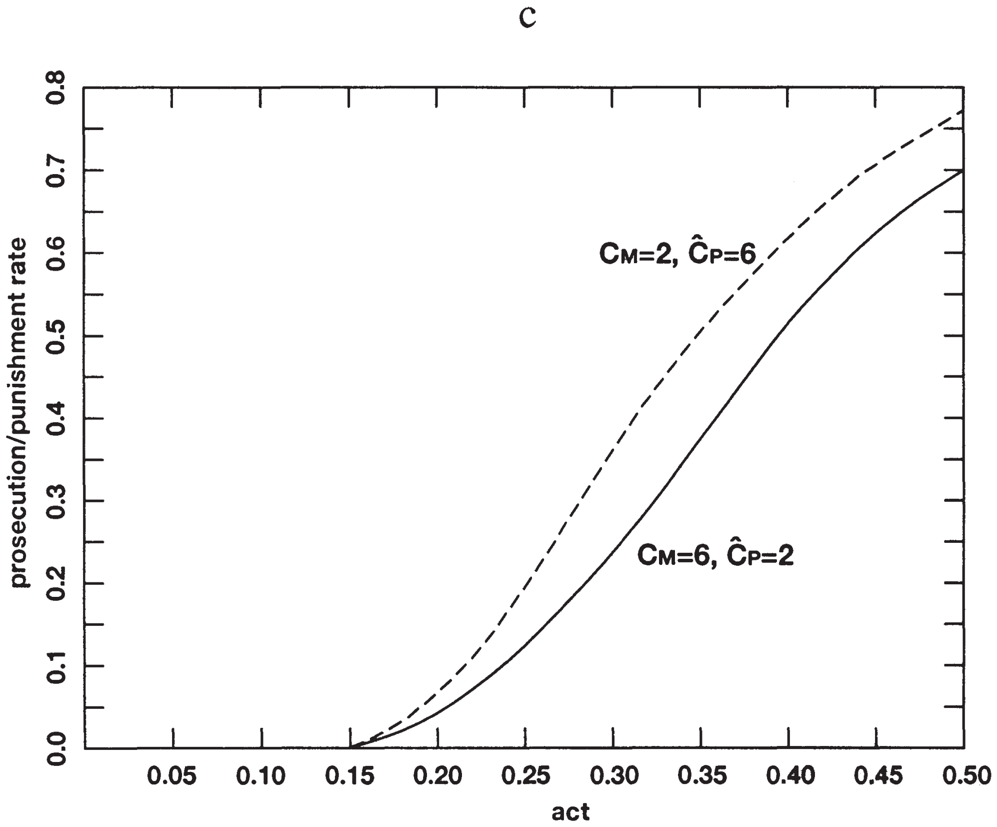

FIG. 4.—*Continued*

Appendix

Proof of the Lemma

In the text, we proved that a schedule a_t can be implemented only if a_t is nondecreasing. To show that (9) is also necessary, note that, by Mirrlees (1986, lemma 6.1), the indirect utility function V satisfies the integral equation corresponding to (6),

$$V(t) = V(0) + \int_0^t b(a_\tau)d\tau,$$

from which (7) follows. The argument in the text then establishes that (9) is necessary.

It remains to show that the two conditions are also sufficient. Given a schedule a_t that is nondecreasing and meets (9), we can set expected penalties according to (10), with $t(a)$ defined as in the lemma. For instance, set $p(a) = 1$, all a, and μ and $f(a)$, all a, according to (12) and (15), respectively. We must show that these expected penalties will implement the schedule a_t.

Recall that $t(a)$ is the highest (supremum) type τ that selects an $a_\tau \leq a$. If a_t is strictly increasing, $t(a_t) = t$, and hence, by (10),

$$tb(a_t) - e(a_t) = \int_0^t b(a_\tau)d\tau \tag{A1}$$

for all t. We contend that (A1) also holds if a_t does not strictly increase everywhere. Suppose, for instance, that $a_s = \hat{a}$, all $s \in [t', t'']$. Then $t(a_s) =$

t'', all $s \in [t', t'']$; hence, by (10),

$$e(a_s) = e(\hat{a}) = t''b(\hat{a}) - \int_0^{t''} b(a_\tau)d\tau = sb(\hat{a}) + (t'' - s)b(\hat{a}) - \int_0^{t''} b(a_\tau)d\tau$$

$$= sb(a_s) - \int_0^{s} b(a_\tau)d\tau,$$

since $a_\tau = \hat{a}$, all $\tau \in [t', t'']$. We now show that (A1) implies that type t will prefer a_t to all other a.

i) First, we show that t prefers a_t to all a_v chosen by some other $v \in [0, T]$. Suppose that $v < t$. Then

$$tb(a_t) - e(a_t) - [vb(a_v) - e(a_v)] = \int_v^t b(a_\tau)d\tau.$$

But, by hypothesis, a_t is nondecreasing and $b(\cdot)$ is increasing; hence

$$\int_v^t b(a_\tau)d\tau \geq (t - v)b(a_v) = tb(a_v) - vb(a_v),$$

and thus

$$tb(a_t) - e(a_t) \geq tb(a_v) - e(a_v).$$

Similarly, when $v > t$, we can also show that t prefers a_t.

ii) Next, we show that t prefers a_t to all $a > a_T$. Note that no type chooses these a. By (10), for $a > a_T$,

$$Tb(a) - e(a) = \int_0^T b(a_\tau)d\tau;$$

hence, by (A1) applied to a_T, we have

$$Tb(a) - e(a) = \int_0^T b(a_\tau)d\tau = Tb(a_T) - e(a_T);$$

that is, type T is indifferent between a_T and any $a > a_T$. For such $a > a_T$, $b(a) > b(a_T)$; hence all $t < T$ will prefer a_T to any $a > a_T$. In part i, we showed that t prefers a_t to a_T; hence he also prefers a_t to all $a > a_T$.

iii) Finally, we show that t also prefers a_t to all $a < a_T$ that are not chosen by some other τ. Since a_t is nondecreasing, such a exist only where the a_t function jumps, say at some t'. If $t' = 0$, then, by (15), $f(a) = 0$, all $a \leq a_{t'}$. Since $b(\cdot)$ is increasing, t' will prefer $a_{t'}$ to any $a < a_{t'}$. Hence, by part i, all higher types will prefer their assigned a_t to any $a < a_{t'}$.

Suppose instead that the jump occurs from a' to a'' at $t' > 0$. All $t < t'$ prefer some a close to a' over a''. Likewise, all $t > t'$ prefer some a close to a'' over a'. Hence, by continuity, type t' must be indifferent between a' and a'', and so also among all $a \in (a', a'')$; that is, $t'b(a) - e(a) = t'b(a'') - e(a'')$, all $a \in (a', a'')$. Since $b(\cdot)$ is increasing, this means that, for all $t > t'$, $tb(a'') - e(a'') \geq tb(a) - e(a)$; that is, higher types will prefer a'' to any $a \in (a', a'')$. Similarly, all $t < t'$ will prefer a' to any $a \in (a', a'')$. Thus, by part i, all types will prefer their assigned a_t to any $a \in (a', a'')$. This completes the proof of the lemma. Q.E.D.

Proof of Proposition 1

We have proved parts a and b and the effect of changes in w in the text; it remains to prove the effect of changes in c_M. We first establish that $c_M + \nu$ cannot fall as c_M increases from c'_M to c''_M. Suppose otherwise. Then, by (14), a_t will fall, from a'_t to a''_t say, at all t, where $a'_t > 0$, and remain unchanged elsewhere. So $\int_0^T b(a_\tau)d\tau$ will fall also. If this causes (9) to be violated, we have a contradiction. Hence a_t must continue to satisfy (9). By (12), monitoring should increase, say from μ' to μ''. Hence

$$\int_0^T [tb(a''_t) - h(a''_t)]g(t)dt - \mu'' c''_M \geq \int_0^T [tb(a'_t) - h(a'_t)]g(t)dt - \mu' c''_M,$$

that is,

$$\int_0^T \{[tb(a''_t) - h(a''_t)] - [tb(a'_t) - h(a'_t)]\}g(t)dt \geq (\mu'' - \mu')c''_M. \qquad \text{(A2)}$$

But, by hypothesis, the schedule a'_t and monitoring μ' were optimal when the monitoring cost was c'_M, so

$$\int_0^T \{[tb(a''_t) - h(a''_t)] - [tb(a'_t) - h(a'_t)]\}g(t)dt \leq (\mu'' - \mu')c'_M.$$

Combining this with (A2), we have $c'_M \geq c''_M$, which is a contradiction.

Therefore, when c_M rises, $c_M + \nu$ either increases or remains unchanged. From (14), the smallest harm subject to enforcement should be the a that maximizes

$$\frac{c_M + \nu}{wg(0)} b(a) - h(a);$$

hence, if c_M increases, the regulators should legalize a wider range of harms. Also, by (14), the schedule a_t rises at all t, implying that $\int_0^T b(a_\tau)d\tau$ will rise; hence, by (12), it will be optimal to reduce monitoring.

Since the optimal a_t rises, the type $t(a)$ corresponding to each harm a falls, implying that the marginal expected penalty $\mu f'(a) = t(a)b'(a)$ also falls. Since the expected penalty for no harm continues to be zero, it also falls for all chosen acts. Q.E.D.

References

Andreoni, James. "Reasonable Doubt and the Optimal Magnitude of Fines: Should the Penalty Fit the Crime?" *Rand J. Econ.* 22 (Autumn 1991): 385–95.

Becker, Gary S. "Crime and Punishment: An Economic Approach." *J.P.E.* 76 (March/April 1968): 169–217.

Cooper, Russell. "On Allocative Distortions in Problems of Self-Selection." *Rand J. Econ.* 15 (Winter 1984): 568–77.

Dickens, William T.; Katz, Lawrence F.; Lang, Kevin; and Summers, Lawrence H. "Employee Crime and the Monitoring Puzzle." *J. Labor Econ.* 7 (July 1989): 331–47.

Friedman, David D. "Reflections on Optimal Punishment, or: Should the Rich Pay Higher Fines?" In *Research in Law and Economics,* vol. 3, edited by Richard O. Zerbe, Jr. Greenwich, Conn.: JAI, 1981.

Friedman, David D., and Sjostrom, William. "Hanged for a Sheep—the Logic of Marginal Deterrence." Manuscript. Chicago: Univ. Chicago, Law School, December 1991.

Landes, William M., and Posner, Richard A. "The Private Enforcement of Law." *J. Legal Studies* 4 (January 1975): 1–46.

Lazear, Edward P. "Salaries and Piece Rates." *J. Bus.* 59 (July 1986): 405–31.

———. "Labor Economics and the Psychology of Organizations." *J. Econ. Perspectives* 5 (Spring 1991): 89–110.

Malik, Arun S. "Avoidance, Screening and Optimum Enforcement." *Rand J. Econ.* 21 (Autumn 1990): 341–53.

Maskin, Eric S., and Riley, John G. "Monopoly with Incomplete Information." *Rand J. Econ.* 15 (Summer 1984): 171–96.

Mirrlees, James A. "The Theory of Optimal Taxation." In *Handbook of Mathematical Economics,* vol. 3, edited by Kenneth J. Arrow and Michael D. Intriligator. Amsterdam: North-Holland, 1986.

Mookherjee, Dilip, and Png, I. P. L. "Monitoring vis-à-vis Investigation in Enforcement of Law." *A.E.R.* 82 (June 1992): 556–65.

Myerson, Roger B. "Optimal Auction Design." *Math. Operations Res.* 6 (February 1981): 58–73.

Polinsky, A. Mitchell, and Shavell, Steven. "The Optimal Use of Fines and Imprisonment." *J. Public Econ.* 24 (June 1984): 89–99.

———. "Enforcement Costs and the Optimal Magnitude and Probability of Fines." *J. Law and Econ.* 35 (April 1992): 133–48.

Posner, Richard A. *Economic Analysis of Law.* 4th ed. Boston: Little, Brown, 1992.

Shavell, Steven. "Specific versus General Enforcement of Law." *J.P.E.* 99 (October 1991): 1088–1108.

———. "A Note on Marginal Deterrence." *Internat. Rev. Law and Econ.* 12 (September 1992): 345–55.

Srinagesh, Padmanabhan; Bradburd, Ralph; and Koo, Hui-wen. "Bidirectional Distortion in Self-Selection Problems." *J. Indus. Econ.* 40 (June 1992): 223–28.

Stigler, George J. "The Optimum Enforcement of Laws." *J.P.E.* 78 (May/June 1970): 526–36.

Wilde, Louis L. "Criminal Choice, Nonmonetary Sanctions, and Marginal Deterrence: A Normative Analysis." *Internat. Rev. Law and Econ.* 12 (September 1992): 333–44.