The Optimal Fine for Risk-Neutral Offenders: A New Approach to the Becker Conundrum

Carl Davidson[a,b], Lawrence W. Martin[a], and John D. Wilson[a,c]

[a] Department of Economics, Michigan State University; East Lansing, MI 48824
[b] GEP, University of Nottingham; Nottingham, UK
[c]CESifo, University of Munich

First Version August 2003
Revised April 2005

Abstract. Gary Becker's classic 1968 paper demonstrates that fines dominate expenditures on detection as a means of controlling illegal activities, because fines represent socially-costless transfers of income. As a result, fines should be set at their maximal levels. Subsequent research has produced several exceptions to this rule, but they involve significant departures from Becker's framework. We use a model that is consistent with this framework (risk-neutral agents and only one type and level of "crime") to demonstrate that it may be optimal to set fines below their maximal levels. The novel feature of our model is that sufficiently high fines interfere with prior commitments made by offenders to compensate other private agents (e.g., debt). In some cases, fines should be low enough to leave offenders with positive assets after honoring all such commitments.

Corresponding Author: Carl Davidson; Department of Economics; Michigan State University; East Lansing, MI 48824; 517.355.7756; 517.432.1068 (Fax); davidso4@msu.edu.

1. Introduction

Gary Becker's classic 1968 paper, "Crime and Punishment: An Economic Approach," created a literature devoted to an investigation of the economics of crime. The typical approach in this literature is to model some sort of regulation or tax that either firms or consumers may try to evade, and then to assume that the government controls illegal activity through random detection and subsequent punishment. The main goal is to characterize the optimal enforcement scheme. Becker's main insight is that fines, a transferable asset with zero social cost, are a superior enforcement mechanism when compared to more likely detection, since detection through audit or inspection is costly. It follows that it always makes sense to substitute a higher fine for a lower probability of detection. As a result, Becker argues that any optimal enforcement scheme would entail a fine set at its maximal level. This finding is sometimes referred to as the "Becker conundrum" because we rarely observe such harsh punishment, though the argument in its favor is clear and compelling.[1]

Over the last 35 years there have been many attempts to conquer the Becker conundrum. For example, it is now well known that harsh fines are not optimal if agents are risk averse (Polinsky and Shavell 1979), because high fines impose an additional risk-bearing cost.[2] In addition, when the illegal activity can take on different gradations, it is optimal to impose only moderate fines on less serious violations, thereby maintaining sufficient marginal incentives to deter more serious offenses. Other approaches concern the optimal treatment of self-reported violations,[3] the structure of the criminal

[1] In a recent survey Polinsky and Shavell (2000) provide a proof that the optimal fine is set at its upper limit when offenders are risk-neutral. Comparing this result with actual practice, they argue for higher fines. "Substantial enforcement costs could be saved without sacrificing deterrence by reducing enforcement effort and simultaneously raising fines."

[2] This approach has been used in the optimal taxation literature to explore the relationship between tax evasion, enforcement and optimal tax policy (see, for example, Sandmo 1981).

[3] In Innes (1999), illegal pollution requires immediate treatment to mitigate its damages. The optimal fine for a violation not reported to the authority is maximal, whereas the optimal fine for a self-reported violation is set lower, reflecting marginal deterrence considerations. If the government is precluded from setting different fines, then the optimal single fine for self-reported and unreported violations is less than maximal. In our models, the government faces no such policy constraints.

justice system,[4] and heterogeneity among offenders.[5] But in the absence of such considerations, existing work implies that the optimal fine for the most serious offenses is maximal when agents are risk-neutral (Shavell 1991, Mookherjee and Png 1992). By maximal fine, we mean that the government seizes all of the convicted offender's available assets.

In this paper, we demonstrate that, even if agents are risk-neutral and there is only one level of illegal activity to be monitored, the fine will still lie below is maximal level, provided monitoring costs are not too high. Indeed, we describe one case where the fine possesses this property regardless of the level of monitoring costs. These results are obtained from a series of models in which firms must decide whether to comply with some sort of government regulation. A distinguishing feature of these models is that there exist agents with claims on the firms' assets. These creditors purchase these claims before they know which firms have chosen to comply with the regulation, and they cannot collect their payments until after the government has assessed and collected fines from convicted non-compliant firms. In such a setting, high fines will interfere with a firm's ability to compensate these creditors; it will declare bankruptcy and provide at most partial compensation.

Our initial model constructs the most favorable case for high fines by demonstrating that fines should bankrupt firms, leading creditors to demand the higher return. Firms in this model differ in the cost of complying with the regulation, and non-compliance creates a negative externality for the economy as a whole. This cost is not known at the time firms choose whether to enter the industry, implying that their entry decisions are made without knowing whether they will comply. Compliance decisions depend on the level of the fine. But once the fine is high enough to send a firm into bankruptcy, any increase in the fine provides no additional incentive to comply with the regulation, because it is effectively being

[4] The structure of the criminal justice system may impose additional costs or otherwise constrain the design of an optimal enforcement system. This literature considers the impact of high fines on the level of resources devoted to defense by defendants (Rubenfeld and Sappington 1987), the amount of resources that criminals devote to avoid detection (Malik 1990), and the likelihood that jurors will vote to convict defendants (Andreoni 1991). See also Acemoglu and Verdier (2000) in which high fines tempt corrupt bureaucrats to accept bribes.

[5] In Bebchuk and Kaplow (1993), offenders differ in their probability of apprehension and in the personal benefits of illegal acts. Reducing enforcement effort can have asymmetric impacts on individual detection probabilities and

paid by the firm's creditors in the form of a reduction in the assets available to repay them. However, fines at this high level remain socially desirable because investors demand a higher return on their capital to compensate them for the risk of bankruptcy, and so firms as a whole bear the burden of a higher fine, turning it into a "Pigouvian tax" on output. In other words, these fines allow the price of output to reflect the negative externalities generated by non-complying firms, plus the increased monitoring costs incurred when output rises. Nevertheless, the fine should not be set at its maximal level, where all of the assets are seized by the government, if these monitoring are sufficiently low.

We devote Section 3 to a series of extensions of the basic model. Naturally, if we allow government to use an actual tax on output, then there is no longer a need for fines high enough to bankrupt firms. But if this is the only change in the model, then this output tax and the fine become perfect substitutes at the margin: what can be achieved with the output tax can also be achieved with a higher fine, implying that bankruptcy-inducing fines remain optimal, but not necessary. But to the extent that tax evasion is a problem, it may be administratively less costly to rely solely on the fine, rather than supplementing the fine with a tax on output.

In a larger departure from the Becker conundrum, Section 3 also includes an extension of the model where the optimal fine not only does not bankrupt firms, but may leave them with positive assets. Unlike the previous models, firms are now given the opportunity to choose their debt levels, and these choices are made prior to the government's monitoring activities. The essential new insight from this model is that firms will use their debt policies to frustrate attempts by the government to use its fine to fully appropriate a firm's assets without also sending it into bankruptcy. Assuming the fine is not too high, the firm will avoid the risk of bankruptcy by lowering its level of debt, in which case the debt will be repaid in full, regardless of whether the firm is fined. But higher levels of debt become cheaper at the margin, because the risk of being fined implies that there is some probability that the firm will declare bankruptcy and not repay all of the debt. In other words, there is a kink in the cost schedule for debt, at

induce an adverse selection effect, where at low enforcement levels, the set of offenders derives lower benefits from their crime than would an equally sized random selection from the population.

the debt level that leaves the firm with zero assets if fined. Consequently, the firm will never choose this debt level. Depending on the level of the fine, it will either choose to lower the effective cost of debt by risking bankruptcy, or it will choose to lower its debt so much that it is left with positive assets after paying the fine. If monitoring costs are sufficiently low, we conclude that the fine should be set low enough to leave the firm with positive assets.

After analyzing this and other extensions of the basic model, we use our concluding remarks in Section 4 to describe two other models that we have investigated elsewhere. In both of these models, we once again find that the fine should not be set at its maximal level.

2. The Optimal Enforcement of Regulations when Firms Borrow

We begin with a simple model of a perfectly competitive industry, in which firms borrow capital from a competitive factor market. These firms face a government-imposed regulation of some sort, and compliance is costly. Some firms may choose to comply with the regulation, whereas others will choose to ignore it and risk detection and punishment. The government's goal is to enforce the regulation in a manner that maximizes social welfare.

To keep things simple, we assume that each risk-neutral firm produces a single unit of the output x. Production of x requires an entrepreneurial input, denoted by e, and a single unit of capital. This capital is provided by investors, who are promised that after all markets clear, they will be repaid the principal of the loan along with interest of r. This principal consists of the unit of capital, which does not depreciate. The cost of complying with the regulation is α, where α is a firm-specific parameter drawn *after the firm enters the market* from a distribution with a continuous distribution function, denoted by $G(\alpha)$. The total cost of production is $e + r + \alpha$ for a firm that has drawn α and decides to comply with the regulation.

Once the firm observes its draw of α, it may attempt to lower its costs by ignoring the regulation. Doing so puts the firm at risk – with probability π, it will be caught and a monetary fine of F will be assessed. This probability of detection is assumed to be the same for all firms.

We now describe the timing of events, leaving the next section to explore an important alternative. In stage one, firms must decide whether to enter the market. If they enter, they hire an entrepreneur and pay him/her *e*, borrow capital, and sign a contract with the investor. In stage two, *α* is revealed, and each active firm makes its compliance decision. In stage three, production occurs and the product market clears. In stage four, the regulatory authority audits firms, detects non-compliance, and assesses fines, which must be paid immediately. Finally, in the last stage, investors are paid. The crucial assumption here is that the government collects fines *before* investors are paid. If the fine is set at a very high level, there may not be sufficient assets available to repay the investor when the firm is detected cheating.

To solve for the equilibrium, we begin with the firms' compliance decision. Suppose that a firm has entered and drawn its value of *α.* As noted above, its costs are $e + r + \alpha$ if it complies with the regulation. If the firm ignores the regulation, then with probability $(1-\pi)$, it gets away with cheating and its costs are $e + r$, but with probability *π*, it gets caught and its costs become $e + F + \min(r, p - F)$ where *p* denotes the output price. Note that in the latter case the investor gets paid fully only if the fine is sufficiently small – that is, if $F \le p - r$. If the fine is higher than this, the firm pays the fine and then turns over all remaining assets to the investor. A risk-neutral firm will be indifferent between compliance and non-compliance if the expected cost associated with non-compliance equals the cost when complying. This occurs at a value of *α,* denoted *α**, which satisfies,

$$\alpha^* = \pi f \,, \tag{1}$$

where f is the "effective fine" paid by a firm, noting that any excess of *F* over $p - r$ is effectively paid by investors through a reduction in the assets they receive from the firm:

$$f \equiv \min(p - r, F) \tag{2}$$

All firms with $\alpha < \alpha^*$ find it optimal to comply with the regulation, while all firms with $\alpha > \alpha^*$ evade it.

Turn next to the firms' entry decisions. Since the firms do not know *α* when they must decide on entry, their expected profits from production are given by:

$$E\Pi(p) = \int_{0}^{\alpha^*} (p - e - r - \alpha) dG(\alpha) + \int_{\alpha^*}^{\infty} (p - e - r - \pi f) dG(\alpha) \qquad (3)$$

For any given π and F, there is a unique p at which these expected profits are zero. At higher values of p, all firms will want to enter, and for lower p, no firm will produce. Setting $E\Pi(p) = 0$ in (3) and rearranging the result gives the following expression for the market-clearing price:

$$p = r + e + \int_{0}^{\alpha^*} \alpha dG(\alpha) + [1 - G(\alpha^*)]\pi f \ . \qquad (4)$$

On the demand side of the product market, there is a single representative consumer with the following quasi-linear utility function: $U(x, p) = E - p + v(x) - h(x_n)$, where E denotes the consumer's endowment of the numeraire good and $h(x_n)$ denotes an external cost created by the output of those firms not complying with the regulation, $x_n = (1-G(\alpha^*))x$. This cost function is assumed to be increasing and convex in x_n. The consumer treats x_n as fixed and chooses x to satisfy the first-order condition, [6]

$$v'(x) = p. \qquad (5)$$

Equations (4) and (5) determine the equilibrium level of x. Since each firm produces one unit of output, the equilibrium number of firms is also equal to x.

Turning to the capital market, we assume that investors obtain capital at the economy-wide rate, or "opportunity cost," of r^* and supply it to the firms. But they cannot observe the firm's α and therefore supply capital without knowing whether the firm will comply with the regulation. If the firm complies with the regulation, or if it successfully evades the regulation, the investors are repaid their capital and earn a return of r. This happens with probability $\delta(\alpha^*) \equiv G(\alpha^*) + (1 - \pi)(1 - G(\alpha^*))$, consisting of the probability $G(\alpha^*)$ that the firm draws a low α and complies with the regulation, and the probability $(1 - \pi)(1 - G(\alpha^*))$ that the firm draws a high α but evades the regulation successfully. Under these circumstances, the investors' profits are $r - r^*$.

[6] We assume that E is large enough that (5) is satisfied for all relevant prices.

If the firm chooses to ignore the regulation and is caught and fined, the profits earned by investors depend on the level of the fine. If the fine is low enough to satisfy $F \le p - r$, the firm earns enough revenue to return the investors' capital and pay the agreed upon return of r. This leads to a profit of $r - r^*$ for the investor. If the firm is in the range where $p - r < F < p + 1$, the application of the fine bankrupts the firm and the regulatory authority seizes a portion of the firm's assets, consisting of revenue p and the unit of capital. In this case, the investors receive the residual, $p - F$ (which may be negative if the fine eats into their principal), and earn profits of $p - F - r^* < r - r^*$. Finally, if the government sets the fine equal to $p + 1$, the authority seizes all of the firm's capital, leaving investors with nothing. In this case, they incur losses equal to $1 + r^*$. Any excess of the fine over this level is not paid. Consequently, there is no loss in generality in restricting F to be less than or equal to $p + 1$, referred to as the "maximal fine."

To summarize, when the firm is fined for non-compliance, the investors' profits are given by $\min\{r - r^*, p - F - r^*\}$. Otherwise, the investor's profits equal $r - r^*$. In equilibrium, investors earn zero expected profits, so if we use $\Gamma(r)$ to denote these profits, we have:

$$E\Gamma(r) = \delta(\alpha^*)(r - r^*) + [1 - \delta(\alpha^*)]\min\{r - r^*, p - F - r^*\} = 0\,. \qquad (6)$$

Given π, p, F and r^*, (6) determines the equilibrium value for r. It is important to note here that any fine above $p - r$ is passed on to all firms, regardless of whether they comply, in the form of a required return r above r^*. This increase in r allows investors to continue to receive an expected return equal to r^*, as implied by (6). The condition for the break-even price, given by (4), may therefore be written as follows:

$$p = r^* + e + \int_0^{\alpha^*} \alpha dG(\alpha) + [1 - G(\alpha^*)]\pi F\,. \qquad (7)$$

We now turn to the government's problem of optimal enforcement of the regulation. The government imposes this regulation because non-compliance generates the external cost, $h(x_n)$. In addition to h, the government must also be concerned about resources that it devotes to enforcement of the regulation. This cost is given by $p_a\pi x$, where p_a denotes the cost of performing each inspection and πx is the total number of inspections carried out. Social welfare is then

$$W = v(x) - h(x_n) - x\{r^* + e + \int_0^{\alpha^*} \alpha dG(\alpha)\} - p_a \pi x . \tag{8}$$

The government's objective is to choose π and F to maximize W, subject to the market equilibrium conditions given above. We assume that lump-sum transfers are available to balance the government budget.

Our main result is that the government chooses a fine that is high enough to cause bankruptcy, but excess liabilities of a bankrupt firm go to zero as the cost of inspection (p_a) goes to zero. These results are contained in the following proposition:

Proposition 1**:** *For $p_a > 0$, any firm that pays the fine is left without profits. Moreover, the fine takes at least some of the firm's capital; that is, $F > p - r$. As p_a goes to zero, F converges to $p - r$.*

Proof: To prove this result, note first that if F were less than $p - r$ (implying $r = r^*$), then both p and the level of compliance, α^*, would depend only on the expected fine, πF, and so a reduction in π and an increase in F that kept πF unchanged would lower audit costs without changing compliance or consumer demand for the good. Thus, the initial F could not have been optimal. This, of course, is the standard Becker argument in favor of high fines.

Thus, we may restrict F to values at least as great as $p - r$. Then the effective fine is $f = p - r$, and (4) shows that this effective fine satisfies

$$f[1 - (1 - G(\alpha^*))\pi] = e + \int_0^{\alpha^*} \alpha dG(\alpha) \tag{9}$$

Thus, the effective fine rises with π but is independent of the statutory fine over those levels where $F \geq p - r$. We may therefore write the compliance rate as a function of π alone: $\alpha^*(\pi)$.

With these properties in mind, let us work with π and x as the control variables for the government's optimization problem. The fine that supports this x is given by the function, $F(x,\pi)$, as defined by equilibrium conditions (1)-(6). Since this fine does not directly enter the welfare function, given by (8), our first-order conditions for x and π will contain no derivatives of this function. One important qualification to this methodology is that it must be checked that the first-order conditions do not imply a fine greater than its maximal level, $p + 1$, since then the government will be unable to support the implied value of x. But we shall see that this condition is satisfied for sufficiently low auditing costs, and for higher costs, an optimal fine equal to p + 1 would satisfy the proposition.

To obtain the first-order condition for x, differentiate (8) with respect to x and use the consumer's optimality condition, $v'(x) = p$, with p given by (7), to obtain

$$\pi F = h'(x_n) + \frac{p_a \pi}{1 - G(\alpha^*)}. \tag{10}$$

The first-order condition for the audit rate is

$$p_a = [h'(x_n) - \pi f]\frac{\partial \alpha^*}{\partial \pi} g(\alpha^*), \tag{11}$$

where $g(\alpha)$ is the density function for α. In the case where $F = f$, these conditions hold if and only if $p_a = 0$. Thus, it is optimal to set $F = p - r$ if and only if there are no auditing costs. In this case, only the expected fine, πF, matters, and the components of this expected fine, π and F, are indeterminate. For positive auditing costs, substitute from (10) for f in (11), to obtain

$$p_a = [\frac{-p_a \pi}{1 - G(\alpha^*)} + \pi(F - f)]\frac{\partial \alpha^*}{\partial \pi} g(\alpha^*). \tag{12}$$

This condition requires that F exceed f, but the difference goes to zero as auditing costs go to zero. #

To understand this result, observe first that that (10) gives us the first-best fine, achievable in the case of costless enforcement: $\pi F = h'(x_n)$. In words, we have the usual Pigouvian prescription for dealing with externalities, adjusted to account for uncertainty: each unit of the good should be taxed at an expected rate, $\pi F(1 - G)$, equal to the expected marginal externality costs that it creates, $h'(1-G)$. This first-best policy accomplishes two goals: (1) the price of the good correctly reflects the marginal externality, so that consumers demand the efficient amount; and (2) firms efficiently choose between compliance and non-compliance, based on a comparison of the social costs and benefits of compliance. Concerning the latter goal, equation (1) gives $\alpha^* = h'(x_n)$ under the first-best policy, implying that firms with $\alpha < h'(x_n)$ chose to comply, and those with $\alpha > h'(x_n)$ do not comply. With free entry, the fine leaves firms with zero expected profits.

If enforcement is costly, then the expected fine, *πF(1-G)*, should still be equated to an "expanded" marginal externality, which consists of both the term *h'(1 – G)* and also the cost of additional audits, $p_a\pi$. Thus, *πF* should exceed *h'*, as described by (10). Note that the actual fine is relevant here, not the effective fine, because any excess of *F* over the effective fine *f* is borne by firms in the form of a higher cost of capital: *r* > *r**. In other words, *F* continues to act like a Pigouvian tax on output.

But for the optimal level of costly audits, only the effective fine *f* matters, because the level of *r* is irrelevant to a firm's choice between compliance and non-compliance. The rule for the optimal audit rate now implies that cost of inducing another firm to switch from non-compliance to compliance, $p_a\left(\frac{\partial \alpha^*}{\partial \pi} g(\alpha^*)\right)^{-1}$, should equal the excess of the reduction in the marginal externality, *h'*, over the loss in effective fine revenue, *πf*. It follows that the effective fine, *πf*, should fall short of *h'*, as described by (11). The only way for this condition to hold while *πF* exceeds *h'*, as required by (10), is for *F* to exceed *f*, implying that the fine forces firms into bankruptcy.

To conclude, costly audits cause the violation of the first-best conditions in two ways: the price of output is too low, leading to too much consumption of the externality-producing good; and there is too much non-compliance. Firms caught not complying with the regulation are forced into bankruptcy. But unless auditing costs are sufficiently high, it is not optimal to set the fine at its maximal level, *p* + *1*, where all of the firm's assets are seized by the government.

3. Extensions

This section presents modifications of the model that produce optimal fine policies that do not entail bankruptcy or, in even greater variance with Becker conundrum, actually leave the firm with positive profits. We conclude this section by briefly showing that the results carry over to a model of tax evasion.

A. Output Taxes

We noted previously that raising F above the break-even point, $p - r$, acts like a tax on output, by raising the cost of capital for all firms. It follows that there should be no reason to bankrupt firms if an actual tax on output is available. Assuming the tax is collected at the time output is produced and sold, it leave firms with assets equal to $p - t + 1$ when auditing occurs. But given our free entry assumption, the tax is passed on to consumers in the form of a higher price, that is, $p - t$ is independent of t. Hence a firm's capacity to pay the fine is not altered by the tax. To minimize auditing costs, F should now be raised at least to $p - t - r$, which becomes the effective fine and is determined by replacing p with $p - t$ in (9). With $p - t$ independent of t, the compliance level continues to depend only on π. However, t now enters the determination of the fine that supports x: $F(x, \pi, t)$. In effect, the fine and tax have become perfect substitutes for manipulating consumer demand for the product. In terms of optimality condition (10), with $\pi F + t$ replacing πF we have:

$$\pi F + t = h'(x_n) + \frac{p_a \pi}{1 - G(\alpha^*)}. \tag{13}$$

Any combination of F and t that satisfies (13) is optimal. In particular, it is optimal to raise F beyond the break-even level, $F \geq p - t - r$, while lowering t to keep (13) satisfied. The only instance in which the availability of the tax improves welfare is when the maximum fine, $F = p + 1$, is reached. In this case, the tax is needed to further reduce output.

With these changes, Proposition 1 should be modified by noting that there are a continuum of optimal fines, all of which satisfy $F > p - t \ - r,$ where $p - t$ is independent of the tax. This continuum shrinks to the single fine, $F = p - r$, as p_a goes to zero, since the need for the tax also goes to zero.

But what if the tax is treated as exogenous, and it is set beyond the maximum level that is optimal when coupled with an optimal fine? It is tempting to conclude that the fine should then be reduced below $p - t -$ r. But this cannot be optimal, because the government would then be able to once again lower auditing costs while increasing the fine to maintain the same level of compliance, with no impact on p. Instead, the government's optimal response to an inefficiently high tax rate will be to lower the expected fine, πF, while maintaining the equality $F = p - t - r$.

Finally, observe that the substitutability between the tax and the fine is based on the assumption that collecting the fine is costless. If tax evasion is a problem, however, then it may not be desirable to introduce a separate tax on output. The reason is that auditing firms for tax compliance requires additional resources. This use of resources is easily avoided by substituting a higher fine for the output tax. Thus, the model appears to justify the use of fines (as opposed to taxes) in part as Pigouvian taxes. We have seen, however, that such fines can be far below their maximal levels if auditing costs are not too high.

B. Input Substitutability

One issue that we have not yet considered is the possibility that fines that drive the firm into bankruptcy will distort the capital market. To investigate this issue, let us remove our fixed-coefficients assumption about production and assume instead that each firm produces its unit of output using some combination of the entrepreneurial input, e, and capital, k, as described by the production relation, $1 = h(e,k)$. For now, assume that these inputs are chosen in the initial stage of the game, when the firm decides to enter the market and does not yet know its cost parameter, α. The critical difference between the inputs is that k is financed with debt, leaving the debtors to compete with the government for repayment after production has occurred, whereas e is financed by the firm owner(s) prior to production. The entrepreneurial-input interpretation suggests various forms of labor, but we could alternatively assume that *e* represents inputs that are equity financed. In either case, a higher value of *e* will require a higher excess of *p* over *r* in the future, given the fine, to compensate the firm owners for the opportunity cost of providing *e*. The equity interpretation of *e* suggests that *e* and *k* might be perfect substitutes in some cases, but we maintain the imperfect-substitutes assumption, reflecting the lack of perfect substitution in practice.

Welfare maximization requires the minimization of unit production costs, evaluated at the social cost of capital, r^*. In symbols, $r^*k + e$ is minimized, subject to the constraint $1 = h(e,k)$. Let *(k*, e*)* denote this cost-minimizing input combination, and suppose that the government attempts to set the fine at a value F^* equal to $p - rk^*$. In other words, investors are paid in full if the firm is fined, but the firm is left with no assets. Realizing that they are paid in full regardless of the fine, investors demand the return $r = r^*$. However, any single firm now recognizes that increasing its *k* beyond k^* will require no additional payments to investors if it is fined, because it will then be bankrupt. Consequently, the firm faces a marginal cost of capital that is below r^* for increases in *k* from k^*, but equal to r^* for decreases in *k.*[7]

As illustrated by the solid lines going through point *a* in Figure 1, the firm's isocost curves over *e* and *k* are kinked at the point (k^*, e^*), implying that this input combination cannot minimize costs. Whereas (k^*, e^*) has been defined so that reductions in *k* raise costs, each firm has an incentive to raise *k*, placing it at the risk of bankruptcy. The only way to eliminate this incentive is to lower the fine, thereby increasing the *k* at which bankruptcy is a possibility. The dotted isocost curve in Figure 1 illustrates the case where the fine is lowered to a level, F^{**}, which makes firms indifferent between k^* and a higher level, k^{**}, under which the fine bankrupts firms. With $F^{**} < p - rk^*$, this lower fine leaves firms with positive assets if they choose k^*.

[7] We are assuming here that investors are not able to monitor a firm's financial structure. This assumption enables us to treat the firm as a competitive price-taker on the capital market, facing an infinitely-elastic supply of capital at the return r*. We could specify some type of imperfect monitoring of the financial structure, but the imperfectness would still leave the firm with a concave capital cost schedule, and sufficient concavity again yields the results reported here.

If audit costs are sufficiently low, then intuition suggests that *F* should be set low enough to leave firms with positive assets in equilibrium. The basic idea is that raising the fine enough to bankrupt firms distorts capital markets: firms face an artificially low price of capital, since they realize they will not have to repay investors in full in the event of bankruptcy, and as a result, they employ too much capital relative to other inputs. On the other hand, lowering the fine enough to prevent bankruptcy requires a higher audit rate to maintain the expected fine at a given level. If audit costs are sufficiently low, however, then the first consideration should dominate the second. We now prove that this is indeed the case.

Proposition 2: *Allowing substitutability between e and k implies that the optimal fine F is strictly below p – k r in equilibrium, if inspection costs are sufficiently low.*

Proof: We begin by allowing the government to directly control the choice of *k* and *e*, and then we transfer this control to firms. With *k* and *e* now subject to choice, the social welfare function, previously given by (8), takes the form

$$W(k,e,\pi,x) = v(x) - h(x_n) - x\{c(k,e) + \int_0^{\alpha^*} \alpha dG(\alpha)\} - p_a \pi x \tag{14}$$

where $c(k,e)$ represents unit social cost,

$$c(k, e) = r^*k + e. \tag{15}$$

With *k* and *e* initially subject to government control, our previous first-order conditions for *x* and *π* remain valid, again implying that as p_a goes to zero, the optimal *F* converges to the level that wipes out all of a firms assets but does not send it into bankruptcy (Proposition 1). In symbols, we can say that *F* converges to $F = p - rk^*$, where $r = r^*$ and (k^*, e^*) is the optimal input combination, which minimizes social cost. Note, too, that if we reduce *F* below this level, while raising *π* to keep the expected fine, *πF*, unchanged, neither *x* nor α^* change, and so differentiation of the welfare function gives:

$$\left.\frac{\partial W}{\partial \pi}\right|_{\pi F = constant} = p_a \textit{ if } p - rk > 0. \tag{16}$$

If $p_a = 0$, then first-order conditions (10) and (11) determine the optimal πF, and this expected fine can be supported by any combination of π and F satisfying $F \leq p - rk^*$. In Figure 2, the solid line measures welfare as F changes, holding fixed the expected fine, k, and e at their optimal values. The flat portion reflects condition (16) for $p_a = 0$, whereas the downward-sloping portion reflects the violation of optimality condition (12) for $F > F^* = p - rk^*$.

Continuing with the case where $p_a = 0$, suppose now that firms are given control of k and e. Since the government had the option of replicating the firms' choices of k and e, handing over these choices to firms cannot raise welfare at any given F and π. However, we can go further and say that the firms and government will choose the same k and e at values of F low enough to eliminate bankruptcy. In this case, firms face the social cost of capital, r^*, for marginal investments and therefore choose k and e to minimize social cost, $c(k,e)$. It follows that the choice of k and e may be transferred to firms without reducing the maximum level of welfare, and this maximum is still obtained at values of F that are low enough to eliminate bankruptcy. The only difference is that these values of F must now leave the firm with positive assets. As previously illustrated in Figure 1, a firm will never choose a k where the fine takes away all of its assets, that is, where $F = p - rk$, since its isocosts are kinked. Rather, the minimum fine, F^{**}, at which the firm is willing to set k low enough to prevent bankruptcy leaves the firm with positive assets, as illustrated by the dashed-line isoquant in Figure 1.[8] But with $p_a = 0$, the required rise in π needed to keep πF optimally set as F is lowered to F^{**} entails no additional inspection cost.

The dashed line in Figure 2 illustrates maximum welfare at different values of F, holding πF fixed at its optimal value. The discontinuous drop in welfare as F rises above F^{**} occurs for two reasons. First, the unit social cost, $c(k,e)$, discontinuously jumps above its minimum level, denoted by c^*. To see this, note that with firms now risking bankruptcy when F is slightly above F^{**},[9] the cost of an additional unit of capital to the firm is $\delta(\alpha^*)r$, where r is again the return paid to firms in the absence of

[8] We make the usual assumption that whenever an agent is indifferent between two strategies, the government is able to induce it to implement the socially-preferred strategy (k^* in this case).

bankruptcy. As determined by (6), *r* guarantees firms an *average* return on capital equal to *r**, which includes the partial compensation paid to investors by bankrupt firms after they pay the fine. If $\delta(\alpha^*)r = r^*$, then firms would continue to choose *k** at the *F* slightly higher than *F***. But with *k** no different than before and the fine only slightly higher, these firms would be able to fully compensate investors after paying the fine. But then *r* would equal *r**, a contradiction. Hence, we conclude that raising the fine slightly above *F*** lowers the marginal cost of capital to a value $\delta(\alpha^*)r < r^*$, causing firms to raise *k* by some discrete amount above *k**. As a result, social cost, $c(k,e)$, jumps above *c**.

The second reason for the drop in welfare as *F* is increased above *F*** is that the resulting rise in *k* (and drop in *e*), combined with the fall in *π* needed to keep *πF* at the level satisfying first-order condition (10), causes the effective fine, $f = p - rk$ to fall below the actual fine, leading to the violation of first-order condition (11). This follows from condition (9), revised to reflect the endogeneity of *k* and *e*:

$$(p - rk)\big(1 - [1 - G(\alpha^*)]\pi\big) = e + \int_0^{\alpha^*} \alpha dG(\alpha)\,. \tag{17}$$

Now if the fine is raised further, it eventually reaches its maximal level, where investors receive no compensation from firms that are fined. In this case, investors demand a return, *r*, where $\delta(\alpha^*)r = r^*$, implying equality between the private and social marginal costs of investments. But the second source of inefficiency still remains: the effective fine is now substantially below the actual fine, implying that first-order conditions (10) and (11) cannot both hold.

The critical implication of these observations is that, if W^b represents the lowest upper bound of welfare over all fines and audit rates that induce bankruptcy, and if W^a is similarly defined over all fines and audit rates that do not entail bankruptcy, then $W^a > W^b$.[10]

[9] The rise in F and drop in π both contribute to bankruptcy. In particular, the zero-profit condition given by (17) below implies that the lower π causes p – rk to fall, reducing the assets available for paying the fine.

[10] In contrast, $W^a = W^b$ in the absence of input substitutability, since the discontinuity in Figure 2 is eliminated.

If we now allow p_a to be positive, then W^a and W^b will change, but for small enough values of p_a, these changes will clearly be small enough to maintain $W^a > W^b$. Thus, bankruptcy continues to be sub-optimal. The only difference is that welfare will now fall as F declines below F^{**}, because the resulting rise in π needed to hold fixed πF will raise audit costs (see eq. 16). Thus, only F^{**} is optimal. #

C. The Timing of Events.

Throughout this study, we have assumed that potential entrants into the industry are ex ante identical, that is, they do not know their cost parameter, α, when making their entry decisions. This assumption distinguishes our analysis from explanations for low fines based on heterogeneous offenders (see footnote 5). Suppose now, however, that the cost parameter is known to firms, but not to investors.[11] We argue in this subsection that this change does not alter our main insights.

One way to model this situation is to posit an upward-sloping supply curve for firms, with the equilibrium price set where marginal entrants earn zero expected profits. These marginal entrants choose not to comply with the regulation, because their cost parameters lie above the cut-off level for compliance, α^*, which again depends on the probability of an audit and subsequent fine. The supply curve therefore slopes up initially, but then turns horizontal over the range of firms not complying. All of the firms with $\alpha < \alpha^*$ earn positive profits.

In this model, a marginal entrant realizes that its expected borrowing costs fall if the fine is high enough to send firms into bankruptcy. Hence, if the fine F lies above $p - r$ (assuming no output taxes or input substitutability), then a marginal entrant that is audited effectively pays a fine equal to $p - r$, with investors effectively paying the remaining fine. Moreover, the entrant's capital is effectively taxed at the rate $r - r^*$, producing a total tax equal to $p - r^*$ on another unit of output. The condition for the optimal level of output is modified by replacing the expected fine πF in (10) with $\pi(p - r^*)$:

$$\pi(p - r^*) = h'(x_n) + p_a \pi \,. \qquad (18)$$

Once again, the expected marginal tax on *x* equals the marginal externality plus the additional auditing cost from another unit of *x*. But observe that $p - r^*$ is less than *F*, because a portion of the fine is now being borne by infra-marginal firms in the form of a higher *r*.

Once again, this higher *r* has no effect on the compliance decision, since it is paid by all firms. Consequently, the condition for the optimal rate, given by (11), remains unchanged; the effective fine in (11) remains given by $f = p - r$. Combining (11) with (18) yields

$$p_a = [\frac{-p_a\pi}{1-G(\alpha^*)} + \pi(r - r^*)]\frac{\partial\alpha^*}{\partial\pi}g(\alpha^*), \tag{19}$$

where the density function *g* now depends on the number of firms that enter the market. Thus, $r > r^*$, implying that the fine should bankrupt firms. It follows that Proposition 1 continues to hold.

The output tax can also be used to eliminate the role of bankruptcy as before, and Proposition 2 also holds when input substitutability is added to the model: if auditing costs are sufficiently small, the optimal fine should leave firms with positive assets. We next argue, however, that the case against bankruptcy seems even stronger than in the previous model with input substitutability.

If the fine is large enough to bring about bankruptcy, then investors again demand a return $r > r^*$ to compensate for this risk. But infra-marginal firms realize that they will not face the risk of bankruptcy. Thus, the higher return on capital that they must pay now represents a distorting tax on their capital investment. By similar reasoning, firms with cost parameters above α^* now know with certainty that they will not comply with the regulation and therefore risk bankruptcy. Hence, they will effectively pay an expected return that lies below r^*. Thus, some firms face capital subsidies and others face capital taxes, both of which create the usual deadweight losses; and a firm that does not comply with the regulation recognizes that lowering its k to the range where bankruptcy does not occur will replace the marginal subsidy on its capital with a marginal tax, again implying a kinked isocost line (Figure 1). For low auditing costs, it will be desirable to eliminate these deadweight losses by setting the fine low enough

[11] Investors must not be able to monitor a firm's capital investment, since they could then infer a firm's type from its investment level.

to leave the firm with positive assets. However, if auditing costs are sufficiently high, then the government may choose to tolerate the deadweight losses.

4. Concluding Remarks

This paper has demonstrated that the optimal regulation of an activity should often not involve the use of maximal fines. In the fixed coefficient version of our model, the fine should be high enough to bankrupt detected non-compliers, but it may still be low enough to enable the firm's debtors to recoup much of their investments. When there is factor substitution, the fine may leave the firm with positive assets even after investors are fully compensated.

The models in this paper assume competitive markets with perfect information in credit and product markets, except for the creditor and government's uncertainty about whether firms are complying with the regulation. In our working paper, Davidson, Martin and Wilson (2003), we introduce informational asymmetries in product markets by assuming that consumers are not fully informed about the quality of the goods they purchase and therefore insist on product guarantees. In particular, consumers demand that each firm post a bond to be forfeited to customers whenever they receive a low-quality good. This strategy of posting a bond to ensure quality supports an equilibrium similar to the one that emerges in Shapiro's 1982 reputation model. But the model encompasses other forms of product warrantees that are effectively equivalent to positing a bond. In each case, the warrantee is assumed to be socially costly. Consequently, the government implements minimum quality standards in an attempt to reduce the equilibrium size of the bond needed to induce firms to produce high-quality goods. To enforce the minimum quality standard, the government again randomly inspects firms and levies fines on those found to be producing low-quality goods. We demonstrate that high fines increase the cost of quality assurance and that the fine should not be set at its maximal level regardless of inspection costs.

The results reported in this paper also extend to models with tax evasion. In our working paper, we investigate a model where a sufficiently high fine prevents a firm from honoring its commitment to pay workers based on their "effort" (which substitutes for other inputs in the production of output). As a result, workers demand a wage premium to compensate them for the risks associated with high fines. We demonstrate again that for sufficiently low audit costs, it is optimal for the fine to leave the firm some assets, which are used to at least partially compensate workers. Moreover, for sufficiently low audit costs, workers are fully compensated, thereby eliminating the required wage premium. Once again, maximal fines are not necessarily optimal.

In all of these models, high fines make it impossible for the detected firm to honor a commitment to compensate a particular private agent. Of course, the agents anticipate this possibility and market prices settle at a level that provides compensation for the expected value of any losses that result when the high fine is assessed. In some cases, this adjustment in market prices is beneficial, such as when the fine acts as a Pigouvian subsidy. In other cases, the fine should be lowered to allow firms to honor their commitments. In our models, the latter cases become more likely as audit costs fall. Indeed, our first model, extended to allow for input substitutability, produces the result that firms should be left with positive assets after paying the fine and fully compensating investors. In this manner, it is possible to fully eliminate the Becker conundrum.

References

Acemoglu, Daron and Thierry Verdier (2000). "The Choice Between Market Failures and Corruption." *American Economic Review* 90: 194-211.

Andreoni, James (1991). "Reasonable Doubt and the Optimal Magnitude of Fines: Should the Penalty Fit the Crime?" *Rand Journal of Economics* 22(3): 385-95.

Babchuk , L. and L. Kaplow (1993). "Optimal Sanctions and Differences in Individual's Likelihood of Avoiding Detection." *International Review of Law and Economics* 13: 217- 24.

Becker, Gary (1968). "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76: 169-217.

Chander, Parkash and Louis Wilde (1998). "A General Characterization of Optimal Income Tax Enforcement." *Review of Economic Studies* 65(1): 165-83.

Davidson, Carl; Lawrence Martin and John D. Wilson (2003). "The Optimal Fine for Risk-Neutral Offenders: Conquering the Becker Conundrum?" Michigan State University Working Paper.

Innes, Robert (1999). "Remediation and self-reporting in optimal law enforcement." *Journal of Public Economics* 72(3): 379-93.

Kaplow, Louis (1990). "A Note on the Optimal Use of Monetary Sanctions." *Journal of Public Economics* 42: 245-47.

Malik, Arun (1990). "Avoidance, Screening and Optimum Enforcement." *Rand Journal of Economics* 21(3): 341-53.

Mookherjee, Dilip and I.P.L. Png (1994). "Marginal Deterrence in Enforcement of Law." *Journal of Political Economy* 102: 1039-66.

Polinsky, Michael and Steven Shavell (1979). "The Optimal Tradeoff between the Probability and the Magnitude of Fines." *American Economic Review* 69: 880-91.

Polinsky, Michael and Steven Shavell (2000). "The Economic Theory of Public Enforcement of Law." *Journal of Economic Literature* 38: 45-76.

Rubinfeld, Daniel and David Sappington (1987). “Efficient Fines and Standards of Proof in Judicial Proceedings.” *Rand Journal of Economics* 18(2): 308-15.

Sandmo, Agnar (1981). “Tax Evasion, Labor Supply and the Equity-Efficiency Tradeoff.” *Journal of Public Economics* 16: 265-88.

Shapiro, Carl (1983). “Premiums for High Quality Products as Returns to Reputation.” *Quarterly Journal of Economics* 98(4): 659-80.

Shavell, Steven (1992). “A Note on Marginal Deterrence.” *International Review of Law and Economics* 12: 345-55.

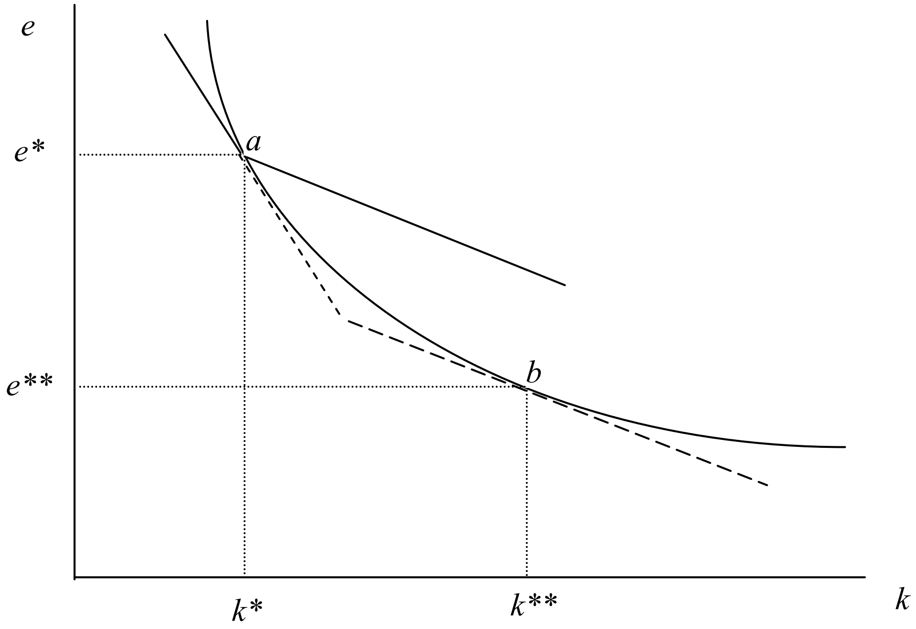

Figure 1

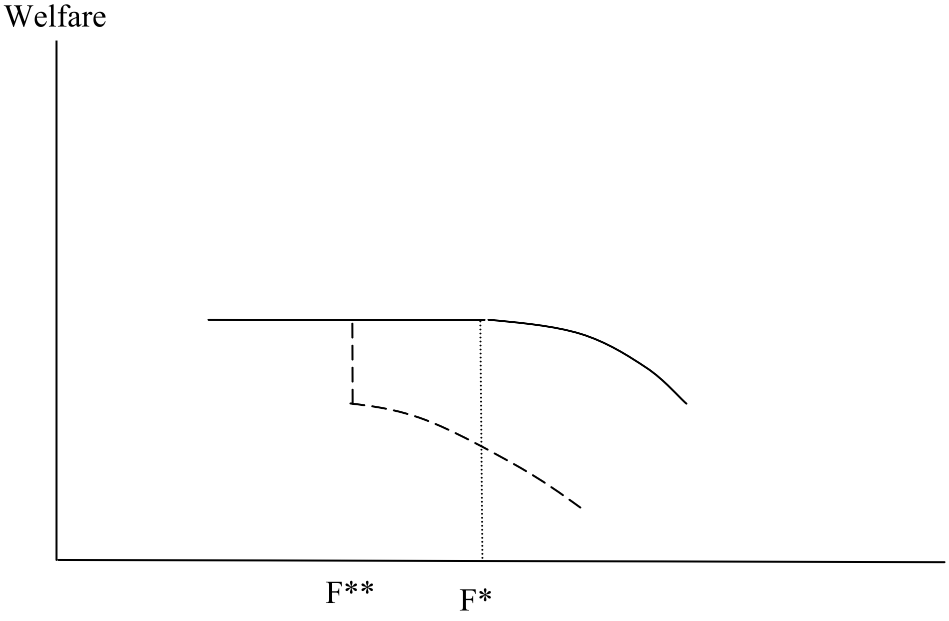

Figure 2