The Market for Illegal Goods: The Case of Drugs

Gary S. Becker and Kevin M. Murphy

University of Chicago and Hoover Institution

Michael Grossman

City University of New York Graduate Center and National Bureau of Economic Research

This paper considers the costs of reducing consumption of a good by making its production illegal and punishing apprehended illegal producers. We use illegal drugs as a prominent example. We show that the more inelastic either demand for or supply of a good is, the greater the increase in social cost from further reducing its production by greater enforcement efforts. So optimal public expenditures on apprehension and conviction of illegal suppliers depend not only on the difference between the social and private values from consumption but also on these elasticities. When demand and supply are not too elastic, it does not pay to enforce any prohibition unless the social value is negative. We also show that a monetary tax could cause a greater reduction in output and increase in price than optimal enforcement against the same good would if it were illegal, even though some producers may go underground to avoid a monetary tax. When enforcement is costly, excise taxes and quantity restrictions are not equivalent.

Our research has been supported by the Robert Wood Johnson Foundation (grant 045566 to the National Opinion Research Center), the Hoover Institution Project on Drugs, the Stigler Center for the Study of the Economy and the State, and the Chicago Initiative on Price Theory. Steve Cicala provided excellent research assistance. Helpful comments were received from Dhaval Dave, Steve Levitt, Ivan Werning, and a referee and at seminars at the University of Chicago and Harvard University. This paper has not undergone the review accorded official National Bureau of Economic Research publications; in particular, it has not been submitted for approval by the board of directors. Any opinions expressed are ours and not those of the Robert Wood Johnson Foundation, the Hoover Institution, NORC, the NBER, the Stigler Center, or the Chicago Initiative on Price Theory.

[*Journal of Political Economy*, 2006, vol. 114, no. 1]

I. Introduction

The effects of excise taxes on prices and outputs have been extensively studied. An equally large literature discusses the normative effects of these taxes as measured by their effects on consumer and producer surplus. This literature claims that quantity reductions are basically equivalent to monetary excise taxes (see Weitzman 1974). However, enforcement of either quantity reductions or excise taxes through apprehension and punishment is largely omitted from these analyses (important exceptions include Glaeser and Shleifer [2001], MacCoun and Reuter [2001], and Miron [2004]).

This paper concentrates on both the positive and normative effects of efforts to reduce quantities by making production illegal and then punishing producers who are apprehended. It compares the effectiveness of such a quantity approach with an excise tax on legal production that punishes only producers who try to avoid the tax through illegal production. We use the supply of and demand for illegal drugs as an important example, a topic of considerable interest in its own right, although our general analysis applies to other efforts to reduce quantity by making production of any good or service illegal, such as prostitution, or restrictions on sales of various goods to minors.

Drugs are a good example because every U.S. president since Richard Nixon has fought a "war" on the production of drugs using police, the Federal Bureau of Investigation, the Central Intelligence Agency, the military, a federal agency (the Drug Enforcement Administration), and the military and police forces of other nations. Despite the wide scope of these efforts—and major additional efforts by other nations—no president or drug "czar" has claimed victory, nor is a victory in sight.

Section II sets out a simple graphical analysis that shows how the elasticity of demand for an illegal good is crucial to understanding the effects of punishment to suppliers. This section considers the interaction between the elasticity of demand and the effects of enforcement and punishment of apprehended suppliers on the overall cost of supplying and consuming that good.

Section III formalizes that analysis systematically and incorporates expenditures by illegal suppliers to avoid detection and punishment. It also derives optimal public expenditures on apprehension and conviction of illegal suppliers by assuming that the government maximizes a welfare function that takes account of differences between the social and private values of consumption of the goods made illegal. Optimal expenditures obviously depend on the extent of the difference between these values, but they also depend crucially on the elasticity of demand for these goods. In particular, when demand is inelastic and enforcement is costless, it does not pay to enforce any prohibition unless the social value is negative and not merely less than the private value.

Section IV generalizes the analysis in Sections II and III to allow producers to be heterogeneous with different cost functions. We show that the negative effect of enforcement against producers of an illegal good on social welfare is greater, not smaller, when the elasticity of supply is smaller. Indeed, supply elasticities enter into the social welfare function more or less symmetrically to demand elasticities. We also show that since enforcement is costly, it is more efficient to direct enforcement efforts toward marginal producers than toward inframarginal producers. By contrast, if the revenue raised by a monetary tax on production is valued, higher monetary taxes should be placed on inframarginal producers because these taxes raise revenue without much effect on output and prices.

Section V compares the effects on costs and output of making all production illegal with the alternative of taxing legal production of the good and punishing underground production only. It shows that a monetary tax on a legal good could cause a greater reduction in output and increase in price than optimal enforcement against production when a good is illegal, even recognizing that some producers may go underground to try to avoid a monetary tax. Indeed, "optimal" quantity with a monetary tax that maximizes social welfare tends to be smaller than the optimal quantity under a policy that prohibits production and punishes illegal producers. This means, in particular, that fighting a war on drugs by legalizing drug use and taxing consumption may be more effective in reducing consumption than continuing to prohibit the legal use of drugs.

Section VI considers whether governments should try to discourage consumption of goods through advertising, as in the "just say no" campaign against drug use. Our analysis implies that such advertising campaigns can be useful against illegal goods that require enforcement expenditures to discourage production. However, they are generally not desirable against legal goods when consumption is discouraged through optimal monetary taxes.

Section VII offers several conclusions, with an emphasis on our results that show the difference between quantity reduction and taxes when enforcement is costly. It emphasizes the importance to the analysis of the elasticity of demand of an illegal product. When demand is inelastic, quantity reductions through enforcement against illegal producers are very costly and can be disastrous.

II. A Graphical Analysis

In an influential article, Weitzman (1974) argues that reducing the consumption of goods either by taxing production with excise taxes or by restricting quantities gives basically equivalent results. However, he ignores the costs involved in enforcing taxes and quantity reductions. Glaeser and Shleifer (2001) bring in enforcement costs in a particular but interesting way. They argue that if the goal is to greatly reduce quantities—as with drugs—it may be easier to enforce quantity reductions than to impose taxes because discovery of quantities is likely to be evidence of illegal production, whereas it may be more difficult to prove that excise taxes were not paid on underground production.

But even Glaeser and Shleifer generally ignore how enforcement operates to reduce quantities when they are made illegal. The drug case illustrates that considerable public resources are usually required to discover illegal production and to punish illegal producers. In essence, the main approach to discourage quantities is to punish producers. When analyzed systematically, this often reverses the conclusion that quantity reductions are cheaper to enforce than monetary taxes.

We first analyze the effects of enforcement expenditures with a simple model of the market for illegal drugs, where the goal is to reduce the quantity of drugs used. The demand for drugs is assumed to depend on the market price of drugs, which is affected by the costs imposed on traffickers through enforcement and punishment, such as confiscation of drugs and imprisonment. The demand for drugs also depends on the costs imposed by the government on drug users.

Assume that drugs are supplied by a competitive drug industry with constant unit costs $c(E)$ that depend on the resources, E, that governments devote to catching smugglers and drug suppliers. In such a competitive market, the transaction price of drugs will equal unit costs, or $c(E)$, and the full price of drugs to consumers, P_e, will equal $c(E) + T$, where T measures the costs imposed on users through reduced convenience or criminal punishments or both. Without a war on drugs, $T = 0$ and $E = 0$, so that $P_e = c(0)$. This free-market equilibrium is illustrated in figure 1 at point f.

With a war on drugs focused on interdiction and the prosecution of drug traffickers, $E > 0$ but $T = 0$. These efforts would raise the street price of drugs and reduce consumption from its free-market level at f to the "war" equilibrium at w, as shown in figure 1.

This figure shows that interdiction and prosecution efforts reduce consumption. In particular, if Δ measures percentage changes, the increase in costs is given by Δc, and $\Delta Q = \epsilon \Delta c$, where $\epsilon < 0$ is the price

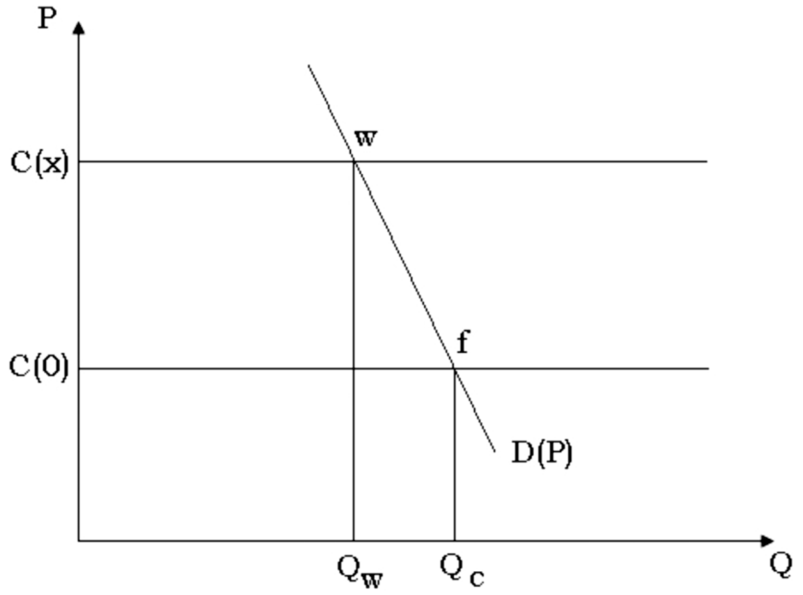

FIG. 1

elasticity of the demand for drugs. The change in expenditures on drugs from making drugs illegal is

$$\Delta R = (1 + \epsilon)\Delta c. \tag{1}$$

When drugs are supplied in a perfectly competitive market with constant unit costs, drug suppliers earn zero profits. Therefore, resources devoted to drug production, smuggling, and distribution will equal the revenues from drug sales in both the free and illegal equilibria. Hence, the change in resources devoted to drug smuggling, including production and distribution, induced by a "war" on drugs will equal the change in consumer expenditures. Therefore, as equation (1) shows, total resources devoted to supplying drugs will rise with a war on drugs when demand for drugs is inelastic ($\epsilon > -1$), and total resources will fall when the demand for drugs is elastic ($\epsilon < -1$).

When the demand for drugs is elastic, more vigorous efforts to fight the war (i.e., increases in E) will reduce the total resources spent by drug traffickers to bring drugs to market. In contrast, and paradoxically, when the demand for drugs is inelastic, total resources spent by drug traffickers will increase as the war increases in severity and consumption falls. With inelastic demand, resources are actually drawn into the drug business as enforcement reduces drug consumption.

The analysis goes through without any fundamental alteration if suppliers of the illegal good act as a monopolist (or full cartel) and if demand has a constant elasticity. For then $Qp = cQ[\epsilon/(1+\epsilon)]$, where cQ is the total cost of production, including costs of punishment and evasion. If ϵ is constant, then the percentage increase in consumer spending (pQ) due to greater enforcement equals the percentage increase in total costs, and so the previous analysis fully applies. However, the evidence that ϵ is about negative one-half contradicts the assumption of a pure monopolistic producer, for such a producer always prices in the elastic section of the demand curve.

III. The Elasticity of Demand and Optimal Enforcement

This section shows how the elasticity of demand determines optimal enforcement to reduce the consumption of specified goods. We assume that governments maximize social welfare that depends on the social rather than consumer evaluation of the utility from consuming these goods. Producers and distributors take privately optimal actions to avoid governmental enforcement efforts. In determining optimal enforcement expenditures, the government takes into account how avoidance activities respond to changes in enforcement expenditures.

We use the following notation throughout this section: Q is the consumption of (e.g.) drugs; P is the price of drugs to consumers; demand is defined as $Q = D(P)$; F is the monetary equivalent of punishment to convicted drug traffickers; production is assumed to be constant returns to scale (CRS) (this is why we measure all cost variables per unit of output); c is the competitive cost of drugs without tax or enforcement, so $c = c(0)$ from above; A is the private expenditures on avoidance of enforcement per unit of output; E is the level of government enforcement per unit of output; and $p(E, A)$ is the probability that a drug trafficker is caught smuggling, with $\partial p/\partial E > 0$ and $\partial p/\partial A < 0$.

We assume that when smugglers are caught, their drugs are confiscated and they are penalized F (per unit of drugs smuggled). With competition and CRS, price will be determined by minimum unit cost. For given levels of E and A, expected unit costs are given by

$$\text{expected unit cost} \equiv u = \frac{c + A + p(E, A)F}{1 - p(E, A)}. \tag{2}$$

Working with the odds ratio of being caught rather than the probability greatly simplifies the analysis. In particular, $\theta(E, A) = p(E, A)/[1 - p(E, A)]$ is this odds ratio, so

$$u = (c + A)(1 + \theta) + \theta F. \tag{3}$$

Expected unit costs are linear in the odds ratio, θ, since it gives the probability of being caught per unit of drugs sold. Expected unit costs are also linear in the penalty for being caught, F.

The competitive price will be equal to the minimum level of unit cost, or

$$P = \min_{A} (c + A)(1 + \theta) + \theta F. \tag{4a}$$

The first-order condition for cost minimization (with respect to A), with E and F taken as given, is

$$-\frac{\partial \theta}{\partial A}(c + A + F) = 1 + \theta. \tag{5}$$

We interpret expenditures on avoidance, A, as including the entire increase in direct costs from operating an illegal enterprise. This would include costs from not being able to use the court system to enforce contracts and costs associated with using less efficient methods of production, transportation, and distribution that have the advantage of being less easily monitored by the government. The competitive price will exceed the costs under a legal environment because of these avoidance costs, A, the loss of drugs due to confiscation, and penalties imposed on those caught.

Hence, the competitive price will equal the minimum expected unit costs, given from equation (4a) as

$$P^*(E) = (c + A^*)[1 + \theta(E, A^*)] + \theta(E, A^*)F, \tag{4b}$$

where A^* is the cost-minimizing level of expenditures. The competitive equilibrium price, given by this equation, exceeds the competitive equilibrium legal price, c, by A (the added cost of underground production); $(c + A)\theta$, the expected value of the drugs confiscated; and θF, the expected costs of punishment.

An increase in punishment to drug offenders, F, raises the cost and lowers the profits of an individual drug producer. For the second-order condition for A^* in equation (5) to be optimal implies that avoidance expenditures increase as F increases. But in competitive equilibrium, a higher F has no effect on expected profits because market price rises by the increase in expected costs due to the higher punishment. In fact, those drug producers and smugglers who manage to avoid apprehension make greater realized profits when punishment increases because those who are caught get punished more, so the increase in market price exceeds the increase in the unit costs of producers who avoid punishment.

The greater profits of producers who avoid punishment, and even the absence of any effect on expected profits of all producers, do not mean that greater punishment has no desired effects. For the higher market price, given by equation (4a), induced by the increase in punishment reduces the use of drugs. The magnitude of this effect on consumption depends on the elasticity of demand: the more inelastic demand is, the smaller this effect is.

The role of the elasticity and the effect on consumption are seen explicitly by calculating the effect of greater enforcement expenditures on the equilibrium price. In particular, by the envelope theorem, we have[1]

$$\frac{dP}{dE} = \frac{\partial\theta}{\partial E}(c + A^* + F) > 0 \tag{6a}$$

and hence

$$\frac{d\ln P}{d\ln E} = \frac{\epsilon_\theta \theta(c + A^* + F)}{P} = \epsilon_\theta\left[\frac{\theta(c + A^* + F)}{P}\right] = \epsilon_\theta\lambda. \tag{6b}$$

Here, $\lambda = \theta(c + A^* + F)/P < 1$ (see eq. [4b]), and $\epsilon_\theta > 0$ is the elasticity of the odds ratio, θ, with respect to E. When we denote the elasticity of demand for drugs by ϵ_d, equation (6b) implies that

$$\frac{d\ln Q}{d\ln E} = \epsilon_d \frac{d\ln P}{d\ln E} = \epsilon_d\epsilon_\theta\lambda < 0. \tag{7}$$

If enforcement is a pure public good, then the costs of enforcement to the government will be independent of the level of drug activity (i.e., $C(E, Q) = C(E)$). On the other hand, if enforcement is a purely private good (with respect to drugs smuggled), an assumption of CRS in production implies that $C(E, Q) = QC(E)$. We adopt a mixture of these two formulations. In addition to these costs, the government has additional costs from punishing those caught. We assume that punishment costs are linear in the number caught and punished (θQ). With a linear combination of all the enforcement cost components,

$$C(Q, E, \theta) = C_1 E + C_2 QE + C_3\theta Q. \tag{8}$$

Equation (8) implies that enforcement costs are linear in the level of enforcement activities, although they could be convex in E without changing the basic results. Enforcement costs also depend on the level

[1] Differentiate eq. (4a) with respect to E and note that in general the optimal value of A will vary as E varies:

$$\frac{dP}{dE} = (c + A^* + F)\frac{d\theta}{dE} + \left[(1 + \theta) + (c + A^* + F)\frac{d\theta}{dA}\right]\frac{dA}{dE}.$$

From the first-order condition for A, the sum of the terms inside the brackets on the right-hand side of the equation for dP/dE is zero.

of drug activity (Q) and the fraction of drug smugglers punished (through θ).

The equilibrium level of enforcement depends on the government's objective. We assume that the government wants to reduce the consumption of goods such as drugs relative to what they would be in a competitive market. We do not model the source of these preferences but assume a "social planner" who may value drug consumption by less than the private willingness to pay of drug users, measured by the price P. If $V(Q)$ is the social value function, then $\partial V/\partial Q \equiv V_q \leq P$, with V_q strictly less than P if there is a perceived externality from drug consumption, and hence drug consumption is socially valued at strictly less than the private willingness to pay. When $V_q < 0$, the negative externality from consumption exceeds the positive utility to consumers.

With these preferences, the government chooses E to maximize the value of consumption minus the sum of production and enforcement costs. Thus it chooses E to solve

$$\max_{E} W = V[Q(E)] - u(E)Q(E) - C\{Q[E], E, \theta[E, A^*(E)]\}. \quad (9)$$

The government incorporates into its decision the privately optimal change in avoidance costs by drug producers and smugglers to any increase in enforcement costs. With the assumption of CRS and perfect competition on the production side, then $u(E)Q(E) = P(E)Q(E)$, and we assume that C is given by equation (8). Thus the planner's problem simplifies to

$$\max_{E} W = V[Q(E)] - P(E)Q(E) - C_1E - C_2Q(E)E$$

$$- C_3\theta[E, A^*(E)]Q(E). \quad (10)$$

The first-order condition is

$$V_q\frac{dQ}{dE} - MR\frac{dQ}{dE} - C_1 - C_2\left[Q + \left(\frac{dQ}{dE}\right)E\right]$$

$$- C_3\left[\theta\frac{dQ}{dE} + Q\left(\frac{\partial\theta}{\partial E} + \frac{\partial\theta}{\partial A}\cdot\frac{dA}{dE}\right)\right] = 0 \rightarrow \quad (11)$$

$$C_1 + C_2\left(Q + E\frac{dQ}{dE}\right) + C_3\left(\theta\frac{dQ}{dE} + Q\frac{d\theta}{dE}\right) = V_q\frac{dQ}{dE} - MR\frac{dQ}{dE}, \quad (12a)$$

where $MR \equiv d(PQ)/dQ$ denotes marginal revenue.

The left-hand side of equation (12a) is the marginal cost of enforcement, including the effects on output and the odds ratio. The right-hand side is the marginal benefit of the reduction in consumption, including the effect on production costs. This equation becomes more revealing if we temporarily assume that marginal enforcement costs are zero. Then the right-hand side of this equation would also equal zero, which simplifies to

$$V_q = MR \equiv P\left(1 + \frac{1}{\epsilon_d}\right) \text{ or } \frac{V_q}{P} = 1 + \frac{1}{\epsilon_d}, \qquad (12b)$$

and V_q/P is the ratio of the social marginal willingness to pay to the private marginal willingness to pay of drug users (measured by price).

If $V_q \geq 0$, so that drug consumption has nonnegative marginal social value, and if demand is inelastic, so that $MR < 0$, equation (12b) implies that optimal enforcement would be zero, and free-market consumption would be the social equilibrium. There is a loss in social utility from reduced consumption since the social value of additional consumption is positive—even if it is less than the private value—whereas production and distribution costs increase as output falls when demand is inelastic.

The conclusion that with positive marginal social willingness to pay—no matter how small—inelastic demand, and punishment to traffickers, the optimal social decision would be to leave the free-market output unchanged does not assume that the government is inefficient or that enforcement of these taxes is costly. Indeed, the conclusion holds in the case we just discussed in which governments are assumed to catch violators easily and with no cost to themselves, but with costs to traffickers. Costs imposed on suppliers bring about the higher price required to reduce consumption. But since marginal revenue is negative when demand is inelastic, total costs would rise along with revenue as price rises and output is reduced as a result of greater enforcement, whereas total social value would fall as output falls if V_q were positive. The optimal social decision is clearly then to do nothing, even if consumption imposes significant external costs on others.

This result differs radically from well-known optimal taxation results with monetary taxes. Then if the monetary tax is costless to implement and if the marginal social value of consumption is less than the price—no matter how small the difference—it is always optimal to reduce output below its free-market level. The reason for the difference is that real production costs fall as output falls with a monetary tax, whereas they rise if demand is inelastic with enforcement of policies that make production illegal. This is just one illustration of the result that enforcement costs can dramatically alter the effect on the total cost of reducing quantities consumed of goods such as drugs.

Even if demand is elastic, it may not be socially optimal to reduce output if consumption of the good has positive marginal social value. For example, if the elasticity is as high as -1.5, equation (12b) shows that it is still optimal to do nothing as long as the ratio of the marginal social to the marginal private value of additional consumption exceeds one-third. It takes very low social values of consumption, or very high demand elasticities, to justify intervention, even with negligible enforcement costs.

Intervention is more likely to be justified when $V_q < 0$: when the negative external effects of consumption exceed the private willingness to pay. If demand is inelastic, marginal revenue is also negative, and equation (12b) shows that a necessary condition to intervene in this market is that marginal social value be less than marginal revenue at the free-market output level.

There are no reliable estimates of the price elasticity of demand for illegal drugs, mainly because data on prices and quantities consumed of illegal goods are scarce. However, estimates for different drugs generally indicate an elasticity of less than one in absolute value, with a central tendency of about one-half (see Cicala 2005), although one or two studies estimate a larger elasticity (see Caulkins 1995; van Ours 1995; Grossman and Chaloupka 1998) and the variability in these estimates is sizable. Moreover, only a few studies of drugs have utilized the theory of rational addiction, which implies that long-run elasticities exceed short-run elasticities for addictive goods (see Becker and Murphy 1988).

Since considerable resources are spent fighting the war on drugs and reducing consumption, the drug war can be considered socially optimal only with a long-run demand elasticity of about negative one-half if the negative social externality of drug use is more than twice the positive value to drug users. Of course, perhaps the true elasticity is much higher, or the war on drugs may be based on interest group power rather than maximization of social welfare.

Punishment to reduce consumption is easier to justify when demand is elastic and hence marginal revenue is positive. If enforcement costs continue to be ignored, total costs of production and distribution must then fall as output is reduced. If $V_q < 0$, social welfare would be maximized by eliminating consumption of that good because costs decline and social value rises as output falls. However, even with elastic demand and negative marginal social value, rising enforcement costs as output falls could lead to an internal equilibrium.

Figure 2 illustrates another case in which it may be optimal to eliminate consumption (ignoring enforcement costs). In this case, demand is assumed to be elastic, and at the free-market equilibrium, V_q is positive and greater than *MR*, but it is less than the free-market price. Marginal revenue is assumed to rise more rapidly than V_q does as output falls, so

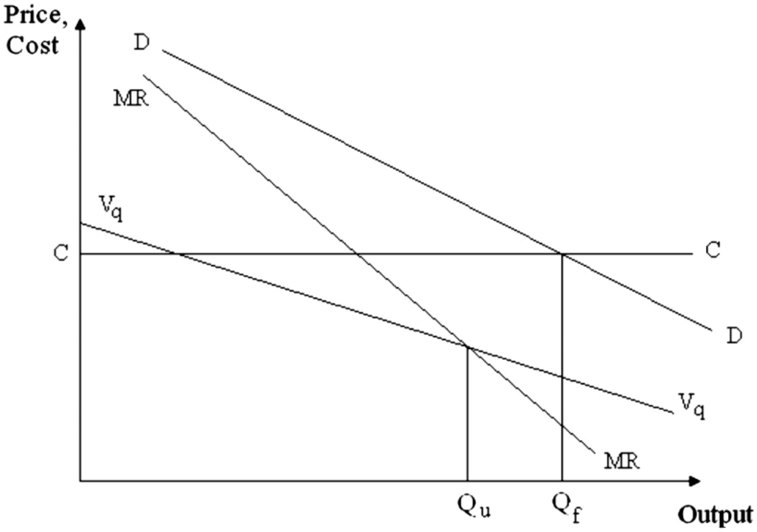

FIG. 2

that they intersect at Q_u. That point would equate *MR* and V_q, but it violates the second-order conditions for a social maximum.

The optimum in this case is to go to one of the corners and either do nothing and remain with the free-market output or fight the war hard enough to eliminate consumption. Which of these extremes is better depends on a comparison of the area between V_q and *MR* to the left of Q_u, with the corresponding area to the right. If the latter is bigger, output remains at the free-market level, even if the social value of consumption at that point were much less than its private value. It would be optimal to remain at the free-market output if reducing output from the free-market level lowers social value by sufficiently more than it lowers production costs.

Equation (12a) incorporates enforcement costs into the first-order conditions for a social maximum. It is interesting that marginal enforcement costs also depend on the elasticity of demand, and they too are greater when demand is more inelastic. To see this, rewrite the left-hand side of equation (12a) as

$$
\begin{aligned}
MC_E &= C_1 + C_2Q + C_2E\frac{dQ}{dE} + C_3\left(\theta\frac{dQ}{dE} + Q\frac{d\theta}{dE}\right)\\
&= C_1 + C_2Q\left(1 + \frac{d\ln Q}{d\ln E}\right) + C_3\left(\theta\frac{dQ}{dE} + Q\frac{d\theta}{dE}\right)\\
&= C_1 + C_2Q\left(1 + \frac{d\ln Q}{d\ln E}\right) + C_3\theta\frac{Q}{E}\left(\frac{d\ln Q}{d\ln E} + \epsilon_{\theta*}\right)\\
&= C_1 + C_2Q(1 + \lambda\epsilon_\theta\epsilon_d) + C_3\theta\frac{Q}{E}\epsilon_{\theta*}\left(1 + \lambda\epsilon_d\frac{\epsilon_\theta}{\epsilon_{\theta*}}\right). \qquad (13)
\end{aligned}
$$

Here $\epsilon_{\theta*}$ is the total elasticity of θ with respect to E, which includes the indirect effect of E on the privately optimal changes in avoidance costs, A, by producers and distributors, that is, since

$$
\frac{d\theta}{dE} = \frac{\partial\theta}{\partial E} + \left(\frac{\partial\theta}{\partial A}\right)\left(\frac{dA}{dE}\right) \rightarrow \epsilon_{\theta*} = \epsilon_\theta + \epsilon_A\frac{d\ln A}{d\ln E}.
$$

Equation (13) shows that marginal enforcement costs are greater, the smaller ϵ_d is in absolute value, because consumption falls more rapidly as enforcement increases when demand is more elastic. Since expenditures on apprehension and punishment depend on output, a slower fall in output with more inelastic demand causes enforcement expenditures to grow more rapidly. Indeed, equation (13) implies that if demand is sufficiently elastic, marginal enforcement costs can be negative when enforcement increases since the drop in the scale of production can more than offset the increased cost per unit.

So the elasticity of demand is key on both the cost and benefit sides of enforcement. When demand is elastic, total industry costs fall as consumption is reduced, and enforcement costs increase more slowly, or they may even fall. Extensive government intervention in this market to reduce output would then be attractive if the marginal social value of consumption is low. In contrast, when demand is inelastic, total production costs rise as consumption falls, and enforcement costs rise more rapidly. With inelastic demand, a war to reduce consumption would be justified only when marginal social value is very negative. Even then, such a war will absorb a lot of resources.

IV. Heterogeneous Taxes and Suppliers

The assumptions made so far of identical firms and of a constant enforcement tax per unit of output have brought out important principles that mainly continue to hold more generally. This section deals briefly with a few novel aspects of optimal enforcement when producers have different costs.

The U.S. experience with the prohibition of alcoholic beverages shows that most producers of these beverages when they were legal exited the industry after prohibition. Legal producers of beer and other alcoholic beverages were replaced by companies that were more willing to deliver beer and liquor to underground illegal retailers, and more skilled at doing so, while evading or bribing the police and courts that enforced prohibition. More generally, suppliers of illegal goods would generally differ from those who would produce and sell the goods when they were illegal.

Presumably, illegal firms would have higher production costs under the contractual and other aspects of the legal and economic environment when production is legal than the firms that produced the goods when they were legal. Otherwise, producers under prohibition would have been the low-cost producers, and they would have dominated the legal industry.

It might seem that our estimate of the loss in welfare of fighting the war on drugs more forcefully is exaggerated by the assumption that the market supply of illegal drugs is completely elastic since costs at the competitive equilibrium would be smaller when supply is inelastic. However, the effect on equilibrium price and quantity of a given enforcement tax per unit is smaller when supply is more inelastic—given the elasticity of demand. This implies that the social effectiveness of expenditures on enforcement to raise price and reduce quantity is smaller when supply is more inelastic. So our assumption of infinitely elastic supply actually lowers the loss in welfare from enforcement to raise costs and price.

To show this simply, let us abstract from all enforcement costs except those imposed on users, which are T per unit of output. Then social welfare is

$$W = V(Q) - C(Q) - TQ. \tag{14}$$

Since $P(Q) = MC(Q) + T$ and $P(Q) - V'(Q) = s$, where s measures the difference between the private and social value of drug consumption,

$$\frac{dW}{dT} = Q\left(-s\frac{d\log Q}{dT} - 1\right). \tag{15}$$

Given the relation between P, MC, and T, it follows that

$$\frac{dQ}{dT} = \frac{1}{P'(Q) - MC'(Q)}, \tag{16}$$

so that

$$\frac{d\log Q}{dT} = \frac{1}{P}\left[\frac{1}{(1/\epsilon_d) - (1/e)}\right], \tag{17}$$

where e is the elasticity of supply. By substitution in the expression in equation (15), we get

$$\frac{dW}{dT} = Q\left\{\frac{s}{P}\left[\frac{1}{(1/e) - (1/\epsilon_d)}\right] - 1\right\}. \tag{18}$$

The last equation on the effect of the enforcement tax on social welfare generalizes our earlier results that assumed $e = \infty$. The formula is symmetric in supply and demand elasticities. If e and $-\epsilon_d \leq 2$ and if $s \leq P$, then enforcement activities at the competitive equilibrium that reduces output must lower social welfare, a substantial weakening of our previous condition that just referred to demand elasticities.

More generally, the lower the supply elasticity, the more likely that greater enforcement activities that raise costs of production, and hence increase market price, would reduce social welfare—given the demand elasticity. As we indicated earlier, the reason for this is that the lower the supply elasticity, the smaller the effect on price and quantity of any given increase in enforcement that raises costs by a fixed amount (T) per unit of production.

If the supply elasticity were less than infinite because some firms are relatively low-cost producers in an illegal environment, the government should be more active in its enforcement against marginal producers and marginal outputs. Any real expenditure on more efficient inframarginal producers and inframarginal units is a waste and serves no efficiency purpose, whereas enforcement against marginal producers helps raise price and thereby induces a reduction in consumption.

With heavier enforcement against marginal producers, the change in producer costs is less than the change in consumer expenditures as the equilibrium price is forced up by enforcement activities. Social costs would then be measured by the smaller rise in producer costs, not by the larger rise in consumer expenditures, as long as the increase in producer rents or profits is considered a transfer from consumers to producers, and not a social cost of the reduction in consumption. However, if no social value were placed on these profits—such as profits to a drug cartel—social cost would still be measured by consumer expenditures, and it would then not be possible to reduce social costs by going after marginal producers.

Of course, it is possible to concentrate on marginal producers only if information is available to enforcers on the costs of different illegal producers. Although the direct information on such costs may be limited, indirect evidence may be considerable since marginal firms tend to be smaller, younger, less profitable, and financially weaker. It would then be efficient to impose higher unit taxes on smaller, younger, and weaker suppliers.

Weaker enforcement against larger producers of drugs is often taken as evidence that these producers bribed and corrupted police and other officials—which may be true. At the same time, our analysis shows that such weaker enforcement may be socially efficient. Government policy should recognize that heavy enforcement against larger and more efficient producers may be a wasteful way to raise price and reduce consumption of drugs, although it may be an effective way to reduce profits to illegal suppliers.

Note the contrast with well-known results on optimal monetary taxation of heterogeneous producers. If tax revenue is highly valued, higher monetary taxes should be extracted from inframarginal producers than from marginal producers because more efficient producers collect profits that can be taxed away, often without major adverse effects on their incentives. In the extreme case of completely inelastic supply, monetary taxes have no effects on incentives or output and produce abundant tax revenue.

V. A Comparison with Monetary Taxes

In this section we show that the equivalence between quantity reductions and excise taxes breaks down completely when quantity is reduced by enforcing a ban on legal production of a good and when enforcement is required to reduce the underground production of a good to escape an excise tax on the good. If we ignore for the moment avoidance and enforcement costs, the social welfare function for monetary taxes that corresponds to the welfare function for enforcement of the prohibition against drugs in equation (9) is

$$W_m = V(Q) - cQ - (1-\delta)\tau Q, \tag{19a}$$

where τ is the monetary tax per unit of output of drugs, and δ gives the value to society per each dollar taxed away from taxpayers. Since in competitive equilibrium $P = c + \tau$, equation (19a) can be rewritten as

$$W_m = V(Q) - cQ - (1-\delta)[P(Q)Q - cQ]. \tag{19b}$$

The first-order condition for Q is

$$V_q = c + (1-\delta)(MR - c), \tag{20a}$$

or

$$\tau = P - V_q + (1 - \delta)\left[P\left(1 + \frac{1}{\epsilon_d}\right) - c\right]. \qquad (20b)$$

If tax receipts are a pure transfer, so that $\delta = 1$, equation (20a) or (20b) gives the classical result that the optimal monetary tax equals the difference between marginal private (measured by P) and marginal social value. With a pure transfer, the elasticity of demand is irrelevant. The optimal monetary tax is then positive if the marginal social value of consumption at the free-market competitive position is less than the competitive price.

The elasticity of demand becomes relevant with net social costs or benefits from the transfer of resources to the government. If government tax receipts are socially valued at less than dollar for dollar ($\delta < 1$) and if demand is inelastic ($\epsilon_d > -1$), the optimal tax would be positive only if the marginal social value of consumption were sufficiently less than the marginal private value. The converse holds if tax revenue is highly valued so that $\delta > 1$. The optimal tax on this good might then be positive, even if demand is inelastic and social value exceeds private value.

Of course, if the monetary tax gets too high, some drug producers might try to avoid the tax by trafficking in the underground economy. Yet an optimal monetary tax on a legal good is still always better than optimal enforcement against an illegal good. The proof assumes that the government can choose optimal punishments for producers who sell in the underground economy and that the demand function for the good is not reduced a lot by making the good illegal.

Let E^* denote the optimal value of enforcement that maximizes the government's welfare function given by equation (10), and recall that this optimal value takes account of avoidance expenditures by producers. Then, from equation (4b), the optimal price is

$$P^* = (c + A^*)[1 + \theta(E^*, A^*)] + \theta(E^*, A^*)F.$$

Assume that enforcement against drug producers who try to avoid the monetary tax by selling in the underground economy is sufficient to raise the unit costs of these producers to the same P^*. If the monetary tax is then set at slightly less than $\tau^* = P^* - c$, firms that produce in the legal sector will be slightly more profitable than illegal underground firms. The latter would be driven out of business or become legal producers. Even if we ignore the revenue from the monetary tax, enforcement costs would then be lower with this monetary tax than with optimal enforcement since few would produce illegally. Indeed, in this case, governments have to incur only the fixed component of enforcement costs, C_1E^*, since in equilibrium no one produces underground.

The government could even enforce an optimal monetary tax that raises market price above the price with optimal enforcement when drugs are illegal. This is sometimes denied with the argument that producers would go underground if monetary taxes were too high. But the logic of the analysis above on deterring underground production shows that this claim is not correct. Whatever the level of the optimal monetary tax, it could be enforced by raising punishment and apprehension sufficiently to make the net price to producers in the illegal sector below the legal price with the optimal monetary tax. Since no one would then produce in the illegal sector, actual enforcement expenditures would still be limited to the fixed component, C_1E^*.

To be sure, the optimal monetary tax would depend on this fixed component of enforcement expenditures. But perhaps the most important implication of this analysis relates to a comparison of optimal monetary taxes and enforcement against illegal goods. If enforcement costs are ignored and if $\delta > 0$, a comparison of the first-order conditions in equations (12b) and (20a) clearly shows that the optimal monetary tax would exceed the optimal "tax" due to enforcement and punishment if demand were inelastic since marginal revenue is then always less than c, unit legal costs of production. The incorporation of enforcement costs only reinforces this conclusion about a higher monetary tax since enforcement costs of cutting illegal output are greater when all production is illegal rather than when some producers go underground to avoid monetary taxes.

If $\delta = 1$ and there are no costs of enforcing the optimal monetary tax, optimal output (Q_f) satisfies $V_q = c$ (see eq. [20a]). When some enforcement costs must be incurred to ensure that no one produces underground, optimal output (Q^*) satisfies

$$(V_q - c)\frac{dQ}{dE} = C_1. \tag{21}$$

Since an increase in E lowers Q, V_q must be less than c. That implies that Q^* exceeds Q_f. Note that optimal legal output is zero when V_q is negative, and there are no enforcement costs. But equation (21) could be satisfied at a positive output level when V_q is negative as long as dQ/dE is sufficiently negative at that output.

Various wars on drugs have been only partially effective in cutting drug use, but the social cost has been large in terms of resources spent, corruption of officials, and imprisonment of many producers, distributors, and drug users. Even some individuals who are not libertarians have called for decriminalization and legalization of drugs because they believe that the gain from these wars has not been worth these costs. Others prefer less radical solutions, including decriminalization of milder drugs, such as marijuana, while preserving the war on more powerful and more addictive substances, such as cocaine.

Our analysis shows, moreover, that using a monetary tax to discourage legal drug production could reduce drug consumption by more than even an efficient war on drugs. The market price of legal drugs with a monetary excise tax could be greater than the price induced by an optimal war on drugs, even when producers could ignore the monetary tax and consider producing in the underground economy. Indeed, the optimal monetary tax would exceed the optimal price due to a war on drugs if the demand for drugs is inelastic—as it appears to be—and if the demand function is unaffected by whether drugs are legal or not; the evidence on this latter assumption is not clear. With these assumptions, the level of consumption that maximizes social welfare would be smaller if drugs were legalized and taxed optimally instead of an optimal reduction in consumption from making production illegal.

The literature on crime and punishment (e.g., Becker 1968) implies that fines are more efficient punishments for illegal activities than imprisonment and other real punishments. Illegal production with fines for those caught could be structured in a way that would make that approach more or less equivalent to a system with taxes on legal production. For example, these systems would be very similar if illegal producers could voluntarily pay a fixed fine per unit produced and if those suppliers who did not were punished sufficiently (perhaps by large enough fines) to discourage underground production.

However, typically, fines to illegal production would not depend linearly on the amount produced, and the size of the fines would rise sharply if suppliers continue after being caught and fined in the past. Moreover, contracts between suppliers of illegal goods and others would not be enforced by courts. As a result, firms that are good at avoiding detection and punishment, and at self-enforcing agreements, perhaps with threats and violence, would have the advantage in a system in which supply is illegal and punishment is achieved through fines. Moreover, some illegal suppliers who are caught may be unable to pay large fines; they would be what is called in legal parlance “judgment proof.” They would have to be punished by imprisonment and with other costly ways. For all these reasons, fines on illegal suppliers and taxing the output of legal suppliers are very different systems.

Our focus in this paper has been on goods with negative externalities for which a prohibition or a tax is a potential way of reducing consumption. We have stressed that a prohibition often operates through raising the costs of suppliers and either increasing market prices or raising the full cost of the good to consumers. For goods with positive externalities, consumption can be increased through either a monetary subsidy or in-kind subsidies designed to lower the costs of producers or lower the full price faced by buyers. Examples would include free parking for patrons, subsidies to building and roads to encourage commerce, and so forth.

It is tempting to believe that the same criticisms we make of in-kind taxes apply to subsidies as well, but this is not the case. With prohibitions and in-kind taxes, the government spends resources in order to raise the costs of the good. Thus if it costs the government $0.50 to raise the unit costs of suppliers by $1.50, the social cost per unit rises $2.00. This amount must be more than offset by the gain from reduced consumption in order for the policy to make sense. In contrast, if it costs the government $1.50 in order to reduce the costs of suppliers by $1.00, the reduction in private costs is subtracted from (rather than added to) the government cost. This gives a net cost of only $0.50, which should be compared to the gain from increased output. When the value to consumers approaches the government cost of providing the subsidy, the efficiency of the in-kind subsidy approaches that of the cash subsidy. This does not happen for in-kind taxes. This advantage of in-kind subsidies over in-kind taxes likely accounts for the much greater frequency of in-kind subsidies by both governments and private firms.

VI. Just Say No

Monetary excise tax theory leaves little room for government policies to reduce the demand function for goods that are taxed. If the purpose is to raise revenue, why try to reduce demand that would lower tax revenue? In addition, it is more efficient to cut consumption because of an externality with optimal monetary taxes that also raise revenue than with costly programs that reduce the demand function.

These advantages do not apply to attempts to lower quantities consumed through apprehension and punishment. Expenditures on enforcement could be reduced by successful government efforts to discourage consumption of certain goods. The campaign to “just say no” to drugs is one example of such an attempt to reduce consumption.

Two types of policy instruments can help reduce consumption of goods such as drugs even when only suppliers are punished: enforcement and punishment strategies that reduce consumption by raising the real costs and prices of supplying the goods and expenditures on “education,” “advertising,” and “persuasion” that reduce demand for these goods. If π represents persuasion expenditures, the social value function W in equation (10) would be modified to

$$W = V(Q(E, \pi), \pi) - P(E)Q(E, \pi) - c(\pi).$$

In this equation, $c(\pi)$ is the cost of producing π units of persuasion against consuming Q, and for simplicity we ignore enforcement costs

(C). We allow W to depend directly on π as well as indirectly through π's effect on Q.

The first-order condition for maximizing W with respect to π is

$$-Q_{\pi}(P - V_q) + V_{\pi} = c_{\pi}. \qquad (22)$$

The term on the right-hand side of this equation, $c_{\pi} > 0$, gives the marginal cost of producing π, and the left-hand side gives the marginal benefit of additional π. If persuasion is effective in reducing consumption, then $Q_{\pi} < 0$. Reduction in consumption is desirable if the marginal social value of consumption, V_q, is less than its private value, measured by P. The sign of the term V_{π} is positive or negative as society likes or dislikes the "persuasion." However, persuasion can have social value even if it is disliked because the left-hand side of equation (22) can be positive, even if $V_{\pi} < 0$, if V_q is sufficiently less than P.

What is interesting about the first-order condition for persuasive activities to reduce demand is that these activities may be effective in raising social welfare when enforcement activities are least effective. We have shown that it is socially optimal not to spend resources to reduce consumption of an illegal good if its demand is inelastic and if the marginal social value of its output is positive ($V_q > 0$).

Equation (22) shows, however, that the elasticity of demand has no effect on the effectiveness of persuasive activities to reduce consumption of an illegal good. Therefore, even if demand is inelastic and even if the marginal social value of its consumption were positive, there still could be a strong case for persuasive efforts to reduce consumption of an illegal good. This depends on whether $V_q < P$, that is, whether marginal social value is less than private value. If it is less, persuasion would raise social welfare if it is cheap to produce and if persuasion efforts do not have a large negative social value. Note that $V_q < P$ is the same criterion that determines whether monetary taxes are desirable.

Persuasion may also raise the effectiveness of enforcement expenditures by raising the elasticity of demand. Becker and Murphy (1993) show that advertising tends to raise the elasticity of demand because it tries to target marginal consumers and increase their demands. It is more efficient for governments to try to reduce demand of marginal consumers than that of other consumers since the former are easier to affect because they get little surplus from consuming certain goods. This means that persuasion does not have to reduce their willingness to pay by a lot to discourage them from consuming these goods. Persuasion could be an effective instrument of government policy not only by reducing the demand for illegal goods but also by raising the effectiveness of enforcement through raising the elasticity of demand for these goods.

VII. Conclusions

Our main conclusions can be stated briefly.

The usually accepted equivalence between quantity reductions and excise taxes fails completely when quantity cutbacks are induced by enforcement and punishment. We show that taxes have a major advantage over quantity reductions when either demand for or supply of the product being taxed is not very elastic, and especially when both are inelastic.

So the elasticity of demand (and supply) plays a major role in our analysis of efforts to reduce consumption of goods such as drugs by making them illegal and enforcing that through punishment of suppliers. Enforcement cuts consumption by raising costs of suppliers mainly because they risk imprisonment and other punishments. The increase in costs leads to higher prices, which in turn induces lower consumption. But if demand is inelastic—as the demand for drugs seems to be—then higher prices lead to an increase in total spending on these illegal goods.

If costs of production, including enforcement costs, are constant per unit of output and supply is competitive, the total real costs of production equal total revenue. Greater enforcement that raises prices will then increase social cost. So social cost would be greater, the harder the push to reduce quantity consumed by raising punishment. Indeed if demand and supply are inelastic, or not very strongly elastic, and even if the social value of consumption of a good was positive although much below its private value, it would not pay to try to reduce quantity consumed below free-market levels by making consumption illegal. The reason is simply that the cost of doing so would exceed the gain.

Excise taxes do not have this problem and can be a much more effective way to reduce consumption, whatever the elasticity of demand and supply. To be sure, it is still necessary to discourage production in the underground economy as producers try to avoid paying the excise tax. However, that can be accomplished more cheaply than when all production is illegal because producers then have the option of producing legally and paying the tax. Enforcement only has to raise the cost of producing in the underground economy above the cost of producing legally in order to discourage illegal production.

This analysis in particular helps us understand why the war on drugs has been so difficult to win, why international drug traffickers command resources to corrupt some governments and thwart extensive efforts to stamp out production by the most powerful nation, and why efforts to reduce the supply of drugs lead to violence and greater power to street gangs and drug cartels. To a large extent, the answer lies in the basic theory of enforcement developed in this paper and the great increase in costs of production from punishing suppliers to fight this war. Suppliers who avoid detection make huge profits, which provides them with resources to corrupt officials and gives them incentives even to kill law enforcement officers and competitors.

References

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

Becker, Gary S., and Kevin M. Murphy. 1988. "A Theory of Rational Addiction." *J.P.E.* 96 (August): 675–700.

———. 1993. "A Simple Theory of Advertising as a Good or Bad." *Q.J.E.* 108 (November): 941–64.

Caulkins, Jonathan P. 1995. *Estimating the Elasticities of Demand for Cocaine and Heroin with Data from 21 Cities from the Drug Use Forecasting (DUF) Program, 1987–1991.* Computer file. ICPSR version. Santa Monica, CA: Rand Corp. (producer); Ann Arbor, MI: Inter-university Consortium Pol. and Soc. Res. (distributor).

Cicala, Steven J. 2005. "The Demand for Illicit Drugs: A Meta-analysis of Price Elasticities." Working paper, Univ. Chicago.

Glaeser, Edward L., and Andrei Shleifer. 2001. "A Reason for Quantity Regulation." *A.E.R. Papers and Proc.* 91 (May): 431–35.

Grossman, Michael, and Frank J. Chaloupka. 1998. "The Demand for Cocaine by Young Adults: A Rational Addiction Approach." *J. Health Econ.* 17 (August): 427–74.

MacCoun, Robert J., and Peter Reuter. 2001. *Drug War Heresies: Learning from Other Vices, Times, and Places.* Cambridge: Cambridge Univ. Press.

Miron, Jeffrey A. 2004. *Drug War Crimes: The Consequences of Prohibition.* Oakland, CA: Independent Inst.

van Ours, Jan C. 1995. "The Price Elasticity of Hard Drugs: The Case of Opium in the Dutch East Indies, 1923–1938." *J.P.E.* 103 (April): 261–79.

Weitzman, Martin L. 1974. "Prices *vs.* Quantities." *Rev. Econ. Studies* 41 (October): 477–91.