OPTIMAL DYNAMIC ALLOCATION OF TREATMENT AND ENFORCEMENT IN ILLICIT DRUG CONTROL

GERNOT TRAGLER

Vienna University of Technology, Department of Operations Research and Systems Theory, Argentinierstrasse 8/1192, A-1040, Vienna, Austria, tragler@e119ws1.tuwien.ac.at

JONATHAN P. CAULKINS

Carnegie Mellon University, H. John Heinz III School of Public Policy and Management; and RAND, 201 North Craig St., Suite 102, Pittsburgh, PA 15213, caulkins@rand.org

GUSTAV FEICHTINGER

Vienna University of Technology, Department of Operations Research and Systems Theory, Argentinierstrasse 8/1192, A-1040, Vienna, Austria, gustav@e119ws1.tuwien.ac.at

(Received June 1998; revisions received March 1999, August 1999; accepted November 1999)

There has been considerable debate about what share of drug control resources should be allocated to treatment vs. enforcement. Most of the debate has presumed that there is one answer to that question, but it seems plausible that the mix of interventions should vary as the size of the problem changes. We formulate the choice between treatment and enforcement as an optimal control problem and reach the following conclusions. If initiation into drug use is an increasing function of the current number of users and control begins early, then it is optimal to use very large amounts of both enforcement and treatment to cut short the epidemic. Otherwise the optimal policy is not to stop the growth of the epidemic, but rather to moderate it. Initially this should be done primarily with enforcement. Over time, enforcement spending should increase, but not nearly so fast as treatment spending. Hence, treatment should receive a larger share of control resources when a drug problem is mature than when it is first growing. If initiation rates subsequently decline, enforcement's budget share should drop further in the ensuing declining stage of the epidemic.

Illicit drugs pose serious challenges for societies around the world. A variety of drug control interventions are available, and an energetic debate has emerged concerning what roles the different interventions should play, with particular focus on the relative merits of treatment and enforcement. Various studies have found that treatment is cost-effective and, in particular, is more effective than enforcement (Gerstein et al. 1994, Rydell et al. 1996, Caulkins et al. 1997). Some (e.g., Crane et al. 1997) have attempted to argue the opposite, and enforcement has consistently been funded more generously than treatment, particularly in the United States. The academic debate has been paralleled among policy makers; e.g., one of the Clinton Administration's early initiatives was a $360 million per year expansion in federal treatment funding (ONDCP 1994), but that was largely blocked by Congress which favored "get tough" approaches.

The goal of this paper is not to argue that enforcement is better than treatment or vice versa, but to explore the possibility that the relative budget shares of these two programs should vary as a drug epidemic evolves. Relatively little cocaine was consumed in the U.S. in the 1960s and early 1970s, but use exploded from the mid-1970s through the early 1980s, and has been fairly stable at much higher levels since then (Everingham and Rydell 1994). The pattern in Europe, as indicated by seizures, appears to be roughly similar but substantially delayed (Farrell et al. 1996). It would be surprising if the optimal policy were to rigidly hold to a particular allocation through such times of rapid change. Since the last century has witnessed recurring waves of drug use (Musto 1987), it seems prudent to prepare for the next wave by asking not only "which program is more effective?" but also "how should the allocation of resources across programs vary over the course of a drug epidemic?"

We try to address this question by applying optimal control theory to a simple model of drug use and drug control that is introduced in the next section. We parameterize that model with data from the recent U.S. cocaine epidemic, even though that epidemic has stabilized, simply because the data for that epidemic are much better than are those for other drugs or other times. In §2 and §3, we analyze two variants of the model, with and without budget constraints, respectively. Section 4 considers how varying assumptions about the nature of initiation into drug use affect the results, and §5 concludes. Mathematical details concerning the analysis are deferred to the Appendix.

Subject classifications: Dynamic programming/optimal control: applications to drug policy. Judicial/legal: crime and drug policy.
Area of review: Services and Military.

This paper is not a final treatise on optimal dynamic drug control for at least two reasons. First, our model is very simple; e.g., it uses just one state variable to reflect drug use. Second, the controls are restricted to treatment and price-raising enforcement. A parallel effort compares treatment and prevention (Behrens et al. 1997) in a model that differentiates between light and heavy users. Ideally one would formulate one model that encompasses all three control programs and different types of users, but having both price effects (necessary when modeling enforcement) and multiple types of users (important in modeling prevention) complicates the analysis. Furthermore, even the restricted model produces important insights for how drug policy should be pursued and evolve over time.

The basic conclusions of this model are that, if one initiates control early, when there are relatively few users, and the problem is truly an epidemic in the sense that initiation into drug use is driven by contact with current users, then one should apply both enforcement and treatment very aggressively to short circuit the epidemic spread. Otherwise the optimal policy is not to stop the growth of the epidemic, but rather to moderate it. Initially this should be done primarily with enforcement, to keep prices high and suppress initiation to the extent possible. Over time, enforcement spending should increase, but not nearly so fast as treatment spending. Hence, treatment should receive a larger share of control resources when a drug problem is mature than when it is first growing. If initiation rates subsequently drop, e.g., because the drug develops a negative reputation, then treatment funding should be reduced as the problem shrinks but enforcement should be cut even more aggressively.

The model generates a variety of other insights, including: (1) detecting the onset of a drug epidemic quickly is valuable because total costs are much lower if control begins early; (2) people who perceive drug use to be costly for society should favor greater drug control spending and allocating a greater proportion of that spending to enforcement; and (3) sharp price declines, such as those observed in the 1980s for cocaine in the U.S., do not necessarily imply a policy failure; indeed it can be optimal to have such declines.

All but one of these conclusions are consistent with the findings of the treatment and prevention model of Behrens et al. (1997). Behrens et al. recommend against using treatment in the early stages of an epidemic because the plight of untreated heavy users might discourage initiation. This model suggests that if one can intervene very early and cut short the epidemic growth, aggressive use of both treatment and enforcement is warranted. Only a unified model can determine whether this model's results could be reversed with a more detailed model of initiation and/or whether the Behrens et al. result could be reversed by the presence of enforcement.

Another comment on differences between the models pertains to the nature of the long run equilibrium. Over the very long run, drug use seems to be cyclic, with periods of epidemic growth followed by a plateau, and slow decline from that plateau. The two-state model (Behrens et al. 1997) studies such cycles, but one-state models such as this one cannot produce cyclic behavior endogenously (Hartl 1987). It only models directly the epidemic growth and plateau. In §4, we exogenously impose a reduction in initiation to crudely approximate the decline stage.

1. THE MODEL

1.1. Intuition Behind the Model

A key characteristic of drugs such as cocaine is that addiction and tolerance make demand a function of the current level of use. So it is important to differentiate users who will consume at a greater or lesser rate depending on the current price from nonusers who do not participate in the market at all until they undergo a discrete transition (called "initiation") and become users. Practically, the number of nonusers is so large that it is effectively constant, so we simply track the number of drug users over time.

We model initiation in two ways. First, we take initiation as a constant, modulated only by the price. Later we consider the case when initiation is increasing in the current number of users, reflecting the fact that most people start using when a friend or sibling introduces them to the drug. The latter approach is more realistic, but results with constant initiation are easier both to obtain and to explain.

There are clearly different intensities of use, and people do not become addicted the minute they initiate, so more refined state spaces are possible (cf. Behrens et al. 1997). We restrict ourselves to a single-state variable, because we want to include price effects without making the model too complicated.

Over time users quit, many of their own accord, but some with the assistance of treatment. That is, treatment can be seen as augmenting the flow out of the population of users. Unfortunately, not all treated users cease use. Relapse is common. Some types of users are more likely to relapse than others, and the treatment system has some capacity to target interventions at those for whom the prognosis is most favorable. Hence, we follow the lead of Rydell et al. (1996) in assuming that treatment's marginal effectiveness diminishes as its scale increases.

Enforcement is quite different. In the first place, most enforcement energy is directed at dealers, not users. (Many users are arrested, but incarceration—which is the more expensive part of enforcement—is more common for distribution offenses; some dealers are also users, but that incapacitative effect is not the dominant aspect of enforcement.) In the second place, enforcement against black markets does not work primarily by removing people from the population. Incarcerated dealers are easily replaced. Rather, enforcement is believed to act more like a tax, raising risks and, as a result, the costs of distributing drugs, which drives up their price. This "risks and prices" aspect of enforcement encompasses not just domestic enforcement, but also interdiction and source country control operations (Reuter and Kleiman 1986). Finally, whereas diminishing returns means that treatment becomes inefficient if it is too large relative to the target, "enforcement swamping" (Kleiman 1993) implies that enforcement is ineffective if it is too small relative to the size of the market. For any given level of enforcement spending, the larger the market over which that effort is spread, the lower the risk and, hence, the smaller the effect on price.

By focusing on prices we ignore two other, less important, ways enforcement helps to control use. First, retail enforcement can drive up "search time" by reducing availability (Moore 1973, Kleiman 1988). Second, interdiction can occasionally generate temporary conditions of physical scarcity. The first is probably of second-order importance (Caulkins 1998). The second has generated some successes (the French Connection/Turkish Opium ban of the early 1970s and the 1989–1990 cocaine price spike), but suppliers can adapt their routes and methods fairly quickly to restore equilibrium conditions (see Caulkins et al. 1993).

Analysts used to reason that since drugs are addictive, consumption must be relatively unresponsive to price, but four recent empirical studies (reviewed by Caulkins and Reuter 1998) suggest that the price elasticity of demand for cocaine in the U.S. is about -1. That is, if prices go up 1%, consumption will go down by 1%. Some of the reduction occurs in the short run, as current users reduce their consumption. Some accrues in the longer term, as higher prices suppress initiation and promote cessation.

We formulate the objective as minimizing the discounted sum of the costs associated with drug use plus the costs of drug control. Quantity consumed has merits as a general purpose measure of the magnitude of a drug problem (Rydell et al. 1996; Caulkins and Reuter 1997), so we assume the societal costs of drug use are proportional to the quantity consumed, where consumption is given by the number of users times the price-modulated consumption rate.

The final piece of our model pertains to the control spending. We consider both unrestricted control (any nonnegative level of treatment and enforcement spending is feasible) and a restricted model in which total spending must be proportional to the number of users. The latter is a crude way of recognizing that budgeting is often reactive, responding to the size of the problem today. In some cases the unrestricted model calls for enormous levels of spending when there are relatively few users in order to prevent future initiation. While such a proactive, aggressive approach can be optimal, it might not be possible to convince taxpayers to spend a lot of money on a potential future problem that is currently small. Likewise we consider a variant of the restricted problem in which not only the level but also the mix of spending if fixed, not optimized dynamically. That is, the decision maker chooses once and for all time what fraction of drug control spending goes to enforcement. The more restricted the set of controls, the worse the objective function value but the easier it would be to implement the optimal control.

1.2. Mathematical Formulation

If we let $u(t)$ and $v(t)$ denote treatment and enforcement spending, respectively, then the forgoing suggests the formulation:

$$\min J = \int_0^\infty e^{-rt}(\kappa\theta A(t)p(A(t),v(t))^{-\omega} + u(t) + v(t))\,dt,$$
$$u(t), v(t) \geqslant 0, \tag{1}$$

subject to

$$\dot{A}(t) = kp(A(t),v(t))^{-a} - c\beta(A(t),u(t))^{z}A(t)$$
$$-\mu p(A(t),v(t))^{b}A(t), \tag{2}$$

where,

$J =$ the discounted weighted sum of the costs of drug use and control,

$r =$ the time discount rate,

$\kappa =$ the social cost per unit of consumption,

$\theta =$ per capita rate of consumption at baseline prices,

$A(t) =$ the number of users at time t,

$p(A(t),v(t)) =$ the retail price,

$\omega =$ absolute value of the short run price elasticity of demand,

$k =$ constant governing the rate of initiation,

$a =$ absolute value of the elasticity of initiation with respect to price,

$c =$ treatment efficiency proportionality constant,

$\beta(A(t),u(t))^{z} =$ outflow rate due to treatment,

$\mu =$ baseline rate at which users quit without treatment,

$b =$ elasticity of desistance with respect to price.

Modeling consumption as $\theta Ap^{-\omega}$ is consistent with a constant elasticity model of per capita demand.

The state dynamics (Equation (2)) has terms for initiation, outflow due to treatment, and the background rate of desistance. For now, the rate of initiation is just a constant, modulated by price. In §4, we make initiation an increasing function of the current number of users by replacing the initiation constant k by $k_2 * A(t)^{\alpha}$, for a positive constant α. The per capita rate of desistance is assumed to be a constant (μ) modulated by price. High prices suppress initiation and encourage desistance. In the absence of controls, the elasticity of the steady state number of users with respect to price is $-a-b$. The overall, or long-run, elasticity of demand is the sum of the elasticity of demand per capita and the price elasticity of the number of users. Hence, we set $-(a+b+\omega)$ equal to the overall elasticity of demand.

Outflow due to treatment is modeled as being proportional to treatment spending per capita raised to an exponent that reflects diminishing returns, with a small constant in the denominator (δ) that prevents division by zero. That is,

$$\beta(A(t), u(t))^z = \left(\frac{u(t)}{A(t)+\delta}\right)^z. \tag{3}$$

We take our model of enforcement's effect on price from Caulkins et al. (1997):

$$p(A(t), v(t)) = d + e\frac{v(t)}{A(t)+\varepsilon}, \tag{4}$$

where ε is an arbitrarily small constant that avoids division by zero. The parameter d captures the fact that prohibition itself forces suppliers to operate in inefficient ways (what Reuter 1983 calls "structural consequences of product illegality"). Because of enforcement swamping, the marginal effectiveness of enforcement (e) is multiplied by enforcement effort relative to market size ($v(t)/A(t)$), not total enforcement effort.

In the constrained budget variant, we have the additional constraint,

$$u(t) + v(t) = GA(t), \tag{5}$$

for a positive constant G. When the mix of interventions is also constrained, we have $v(t) = fGA(t)$ and $u(t) = (1-f)$ $GA(t)$.

1.3. Parameters

Parameter values were chosen as described in Tragler et al. (1997). In brief, values for a, b, and ω together reflect a belief that the long-term price elasticity of demand is -1 (base case in Caulkins et al. 1997), that the short term elasticity (ω) is half the long-term elasticity (as observed by Saffer and Chaloupka 1995 for cocaine, and Becker et al. 1994 for cigarettes), and that the long-term portion can be divided equally between effects on initiation and exit (as in Rydell and Everingham 1994).

The average annual rate of initiation reported by Johnson et al. (1996) between 1972–1992 was 1,034,571, so we choose k to make initiation be 1,000,000 at the baseline price, which ONDCP (1997) reports was \$106.73 per pure gram in our base year of 1992. ONDCP (1996) reports the average number of users in 1992 and 1993 was 6,486,000, so we choose parameters that make $A = 6{,}500{,}000$ in base conditions. Rydell and Everingham (1994) report total spending on cocaine treatment and enforcement of \$10.4 billion; dividing by 6.5 million users gives $G = \$1{,}600$ per user. They estimate total consumption at 291 metric tons, so we set $\theta = 14.6259$ (since $14.6259 * 0.10673^{-0.5} = 291{,}000{,}000/6{,}500{,}000$ and price is expressed in thousands of dollars).

Rydell and Everingham (1994, p. 38) report cocaine-related health and productivity costs of \$19.68 billion for cocaine in 1992, which is associated with 291 metric tons of consumption, implying an average cost of \$67.6 per gram (in 1992 dollars). These figures do not include crime-related costs, so in light of Miller et al. (1996), we take $\kappa = \$100/\text{gram}$ as our base value for the social cost per gram of consumption.

The price function parameters (d and e) reflect a price of \$106.73 per gram under base case enforcement spending and an elasticity of price with respect to enforcement spending of 0.3636 as in Caulkins et al. (1997).

Following Rydell and Everingham (1994), we assume an average of \$1,700–\$2,000 is spent per admission to treatment and that this provides a 13% chance of ceasing heavy use, over and above baseline exit rates in the absence of treatment. About one-third of users are heavy users (ONDCP 1996), so at baseline rates of spending, about 30% of heavy users received treatment (Rydell et al. 1996); the outflow rate due to treatment was 85,800 users per year, and, thus, $c = 0.04323$. We have no direct measure of the rate of diminishing returns, so we set $z = 0.6$ because that causes treatment's steady-state budget share in the optimally controlled allocation problem to be around one-third. One-third is somewhat more than treatment's actual share in the U.S. in the 1990s, but we are persuaded by studies such as Rydell et al. (1996) that the current fraction is smaller than would be optimal.

The outflow parameter μ was selected to make the outflow rate be 700,000 per year at base prices, which reflects the observed population change (ONDCP 1996) net of initiation and treatment during the recent years of relative stability. The discount factor is set at $r = 0.04$ as in Rydell et al. (1996) and Caulkins et al. (1997).

The values are summarized in Table 1. Two values are given for parameters d, e, k, κ, θ, and μ. The values in brackets are the ones just described. For analytical convenience, we set $\kappa = 1$ and $\theta = 1$, and adjust d, e, k, and μ accordingly, yielding the second set of values for those parameters.

2. THE CONSTRAINED BUDGET PROBLEM

Consider first the case in which controls are constrained so that $u(t) + v(t) = GA(t)$. The model cannot be solved analytically, but the numerical solution, guided by the maximum principle as described in the Appendix, is summarized by the phase portrait in Figure 1. (In the sequel, we omit explicit reference to the time argument where there is no danger of confusion.)

The steady state values are given by simultaneously setting the derivatives of the state (A) and control (v) variables equal to zero, yielding the gray curves in Figure 1.

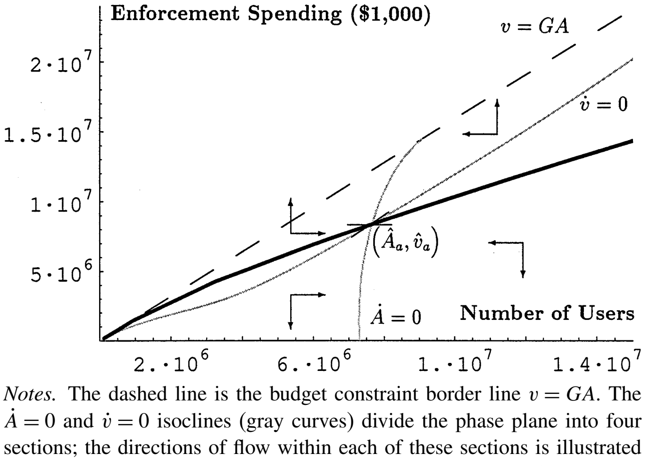
Table 1. Base case parameter values.

Parameter	Value	Description
a	0.25	negative elasticity of initiation with respect to price
b	0.25	elasticity of desistance with respect to price
c	0.04323	treatment efficiency proportionality constant
d	0.03175 [0.06792]	price with minimal enforcement (in thousands of $)
δ	0.001	constant to avoid division by zero
e	0.01241 [0.02655]	enforcement efficiency proportionality constant
ε	0.001	constant to avoid division by zero
G	1.6	control budget (in thousands of dollars) per user
k	472,618 [571,573]	initiation constant (new users per year)
κ	1 [0.1]	social cost per gram consumed (in thousands)
μ	0.22786 [0.18841]	natural outflow rate from use
θ	1 [14.6259]	per capita consumption constant
ω	0.5	negative of short run elasticity of demand
r	0.04	annual discount rate (time preference rate)
z	0.6	$1-z$ reflects treatment's diminishing returns

Using the parameter values from Table 1, the intersection of these curves in the A-v plane is a saddle point equilibrium $(\widehat{A}, \hat{v})$. Every saddle point equilibrium in a two-dimensional phase portrait has a stable manifold which consists of two branches. Locally, these branches are determined by the eigenvector associated with the negative eigenvalue of the Jacobian evaluated at the steady state. This is used to numerically compute all stable manifolds (black curves) which, in optimal control theory, are known to be candidates for the optimal trajectories.

Figure 1 implies that if control begins when the number of users is below its steady state value ($A(0) < \widehat{A}$), the optimal treatment and enforcement rates start low and gradually increase while $A(t)$ converges to $\widehat{A}$. (The opposite holds for initial states above the steady state value.) Furthermore, since the stable manifold is tangent to the budget line, when there are few users, almost all the resources should be allocated to enforcement.

Figure 1. Phase portrait in the A-v-plane for the constrained budget problem.



Notes. The dashed line is the budget constraint border line $v = GA$. The $\dot{A} = 0$ and $\dot{v} = 0$ isoclines (gray curves) divide the phase plane into four sections; the directions of flow within each of these sections is illustrated by the arrows. The black curves are the stable manifolds of the saddle point equilibrium $(\widehat{A}_a, \hat{v}_a)$ (intersection of the isoclines)

Even if the optimal policy is pursued, the number of users will increase over time toward the equilibrium $(\widehat{A})$. Enforcement spending should also increase, but less than proportionately, so that over time, a larger and larger share of control effort should be allocated to treatment.

It does not appear that U.S. policy followed this pattern. It is difficult to obtain data on drug-specific spending, particularly data which combines federal, state, and local efforts, but the ONDCP (1996, pp. 318–319) reports federal spending on drug control generally (not just on cocaine). Between 1981 and 1990, treatment's share fell from 33.5% to 16.8%, and has remained between about 17% and 20% since.

Our principal objective was to understand how the budget shares for enforcement and treatment should vary over time, but sensitivity analysis generates other insights. For example, the inability to completely eradicate drug use is not primarily due to inadequate control spending. Even spending levels one hundred times greater than the current level of $G = \$1{,}600$ per user are not enough to eradicate drug use with the constant initiation function. As we will see later, if initiation is an increasing function of the current number of users, it will be possible and in some cases optimal to eradicate drug use by expending sufficient resources.

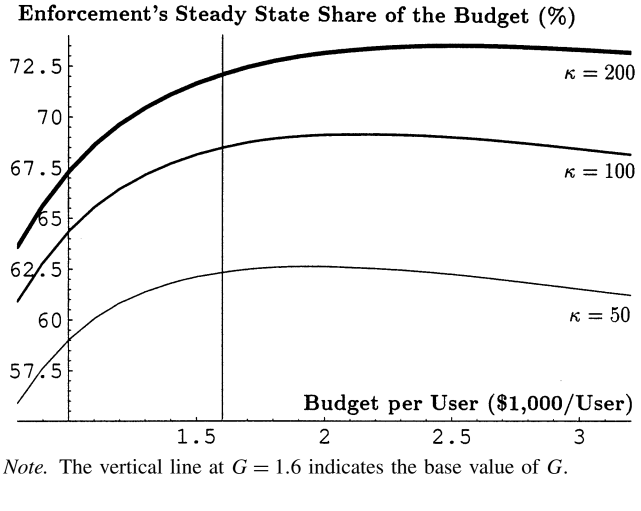
Another question relating G to the number of users is similar to that raised by the so-called "Laffer curve" (see, Varian 1996). Laffer observed that if increasing the marginal tax rate decreases the number of hours people work, it might increase or decrease tax revenues depending on what the current tax rate is. Similarly increasing control spending per user decreases the number of users, so it might or might not increase total spending. However, just as the ironic possibility that the Reagan Administration's tax cuts might lead to smaller deficits failed to materialize, for this model and parameter values, one would have to spend enormously more on drug control per user before total drug control spending began to fall. In particular, total spending is still increasing in G for values of G one hundred times greater than the current level.

Sensitivity of equilibrium values to the parameters is also insightful. (See Tragler et al. 1997 for a detailed discussion.) One result we will use below is that the steady state number of users and levels of control spending scale almost linearly in the initiation proportionality constant in the sense that their elasticities with respect to the initiation constant are very close to unity. The most interesting results in and of themselves, though, pertain to the interaction of the social cost per gram of consumption (κ) and the size of the drug control budget per user (G). Sensitivity with respect to social cost is important not only because of the usual uncertainties, but also because different people may wish to include different outcomes in the objective and/or ascribe different subjective weights to their importance (e.g., different people might assign different values to the social cost of having someone be addicted to a psychoactive substance).

Figure 2 shows what proportion of funding is optimal to allocate to enforcement in steady state as a function of the size of the drug control budget per user (G, horizontal axis) and the social cost per gram of cocaine consumed (base case is $\kappa = \$100$ per gram). The higher the social cost of consumption is perceived to be, the higher should be enforcement's share of the budget. Also, for any given social cost estimate, there seems to be one value of G for which the optimal enforcement share reaches a maximum. That is, if society were willing to spend either more or less per user on drug control, treatment's share of the budget should increase. For the base case parameter values, the current spending level (G) is close to the value that maximizes the share of the budget that should be allocated to enforcement.

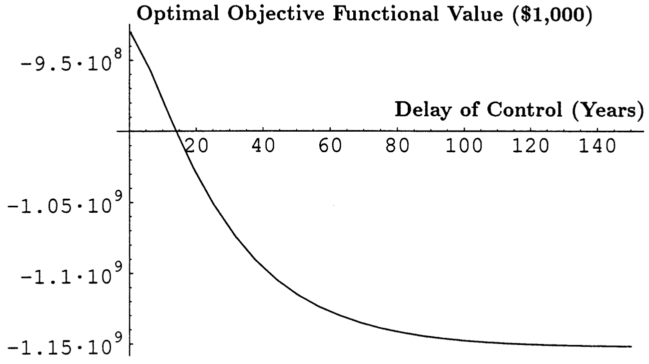
Related insights can be derived from a plot of the optimal utility functional value as a function of these two parameters (see Tragler et al. 1997, for figure). The greater the perceived social cost per unit of use, the more it is optimal to spend per user. With the current parameter values, people who think social costs are relatively low ($50 per gram) should argue for lower control spending (smaller G). People who think social costs are high ($200 per gram) should argue for a substantial increase. People who believe our base estimate ($100 per gram) should be pretty happy with the current level of effort, perhaps favoring a slight expansion. Given the imprecision in our parameter estimates, we do not ascribe great import to these numerical results, but the consistency of their interpretation with the current policy debate is interesting.

Figure 2. Optimal fraction of resources allocated to enforcement in steady state as a function of the drug control budget per user (G) for three levels of social cost per gram consumed ($50, $100, and $200/gram).



Note. The vertical line at $G = 1.6$ indicates the base value of G.

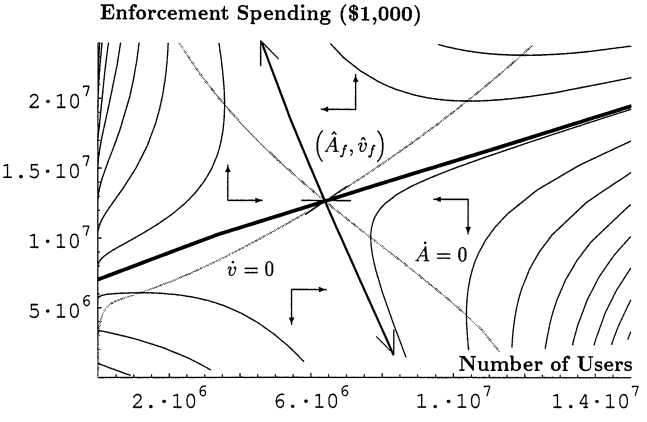
Figure 3. Optimal utility functional value as a function of τ, the delay before control begins.



The qualitative relationship (G^* increasing in κ) holds whether the initial number of users $A(0)$ is small or large, but the optimal intensity of control (G) is greater when the initial number of users is smaller, suggesting that if control begins early, it is particularly valuable. Figure 3 confirms this final insight by plotting the objective functional value vs. how long it takes for control to begin after the point when there are $A = 100{,}000$ users. Beginning control promptly could avert one-quarter of the total cost associated with the drug epidemic, even if the level of control is constrained to be proportional to the number of users (i.e., without using a massive initial intervention to short circuit the epidemic). The reduction in costs is nearly linear over the first 20 years of delay, and actual delays are likely to be no more than 20 years since by that time the number of users is quite large. So it makes sense to think of a cost per year of delay, and that figure is about $5.5 billion per year (net present value of all future costs).

Suppose there is no epidemic at present, but that the expected time to the next epidemic is on the order of 25 years. Then, ignoring the time-value of money, one should be willing to spend up to $5.5 billion divided by 25 years = $220 million per year on a monitoring system that would detect the next epidemic a year earlier than it otherwise would be. This figure dwarfs the current U.S. investment in drug-related data collection and analysis, and in important respects the U.S. data systems are far better than those in most if not all other countries. The model in this paper is more reliable for qualitative not quantitative results and the 25-year figure has no particular basis, so the $220 million

Figure 4. Phase portrait in the A-v-plane for the problem without a budget constraint (cf. Figure 1).



per year figure should be taken with a big grain of salt, but it does underscore the value of information in the context of early intervention in a drug epidemic.

3. THE UNCONSTRAINED MODEL

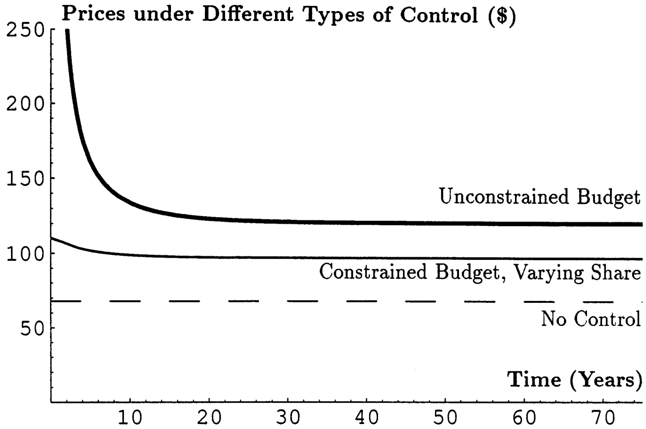
We turn next to the unconstrained problem in which total control spending can be larger or smaller than G times the current number of users when that is desirable. The solution is described in the Appendix and summarized in a phase portrait (Figure 4). As before, the gray curves show the locus of points for which the time derivatives of the state and control variables are zero. Again using the base parameter values from Table 1, the intersection of the isoclines is a saddle point equilibrium $(\widehat{A}, \hat{v})$, and the two stable manifolds (thick black curves) yield the optimal trajectories.

This structure appears to be robust with respect to the parameter values. We varied every parameter between 50% and 150% of its base value and obtained a single steady state that is a saddle point in every case but one. If z, the exponent in the treatment efficiency function, increases to 130% of its base value (i.e., from 0.6 to 0.78) then a second steady state emerges that is an unstable focus. This yields a solution similar to that in Figure 6.

For enforcement spending (v), there is a striking difference between the solutions in Figures 4 and 1. When the budget is not constrained, it is optimal to spend a great deal per user on enforcement when there are relatively few users. Enforcement spending should be roughly linear in the number of users, but with a large intercept.

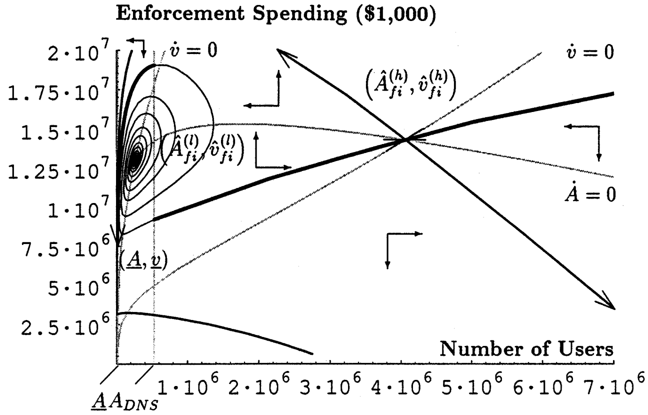
Unlike the constrained budget problem, we cannot simply infer the optimal level of treatment spending (u) from the difference between a budget constraint and the optimal level of enforcement spending (v). It turns out that treatment spending should be very nearly proportional to the number of users with a slope comparable to that for enforcement (see Tragler et al. 1997, Figure 3.13). Thus, as before, treatment's proportionate share of the control budget should increase as the number of users grows toward its equilibrium level.

Figure 5. Time paths of optimal prices with no control ($p_n(t)$), constrained budget ($p_a(t)$), and control free of budget constraints ($p_f(t)$).



Since enforcement spending is roughly linear with a large intercept, total spending per user is initially very large and falls steadily as the number of users increases (decreasing to $2,880 per user in equilibrium). If control starts with $A(0) = 100{,}000$ users, spending per user is very large for about seven years, but falls to within 10% of its equilibrium value within 14 years. This has interesting implications for the optimal price trajectory, namely that it can be optimal to have prices decline precipitously as the number of users grows (See Figure 5). That, in fact, happened in the U.S. during the 1980s. Often that price collapse is thought of as a disaster, and it may have been a disaster, but it is also possible that it was the consequence of an optimal policy.

Figure 6. Phase portrait in the A-v-plane for the unconstrained budget problem with state dependent initiation (cf. Figures 1 and 4).



Note. The two gray vertical lines indicate the lower limit on the number of users, $\underline{A} = 10{,}000$, and the DNS threshold, $A_{\text{DNS}} = 529{,}117$. There are two $\dot{v} = 0$ and one $\dot{A} = 0$ isoclines (gray curves). The optimal trajectories are given by the thick black curves: for initial numbers of users below the DNS point the movement is towards the lower limit steady state $(\underline{A}, \underline{v})$, above A_{DNS} the optimal trajectories lead to the high volume equilibrium $(\widehat{A}_{fi}^{(h)}, \hat{v}_{fi}^{(h)})$; the steady state $(\widehat{A}_{fi}^{(l)}, \hat{v}_{fi}^{(l)})$ is an unstable focus.

Table 2. Equilibrium levels of use and control spending.

Model of control	$\widehat{A}$ (millions)	$\hat{u}$ ($B/yr.)	$\hat{v}$ ($B/yr.)	J^* ($B, NPV)
No control	11.64	$0	$0	$1,157
Constrained budget with $G = \$1{,}600$, fixed share	7.72	$3.42	$8.93	$931
Constrained budget with $G = \$1{,}600$, varying share	7.65	$3.86	$8.39	$913
Constrained budget with $G = \$3{,}310$, fixed share	6.13	$5.80	$14.49	$900
Unconstrained budget	6.40	$5.78	$12.67	$886

Table 2 contrasts the equilibrium approached without a budget constraint to those approached with three variants of the constrained budget problem (fixing enforcement's budget share and allowing it to vary dynamically with $G = \$1{,}600$, and fixing it with $G = \$3{,}310$, the best value of G with a fixed budget share) and with no controls. (*Note: uncontrolled* means no additional effort beyond routine policing; it does not model legalization.) With or without the budget constraint, controls reduce social costs in equilibrium by about one-quarter (starting with $A(0) = 100{,}000$ users). With no budget constraint, equilibrium spending per user is 80% higher than if spending is constrained at current levels ($2,880 per user vs. $G = \$1{,}600$ per user), but the proportion going to enforcement is nearly identical and the objective functional value (J^*) is similar (just 3% lower). The benefits of that additional spending's ability to suppress the number of users by one-sixth are almost all offset by the increased control costs.

Table 2 also reveals that it makes surprisingly little difference whether the proportion going to enforcement is fixed or allowed to vary. The most important thing is to have some control. The second priority is choosing the right amount of control ($G = \$3{,}310$ vs. $G = \$1{,}600$). At that point allowing budget shares to vary only saves a few billion dollars more. On the other hand, choosing the right value of G is hard, in part because of the usual parameter uncertainty, in part because it depends on judgements about social costs, which are inherenly "soft." In contrast, it should be easier (theoretically if not bureaucratically) to vary budget shares at least approximately, as suggested in Figures 1 and 4, because it is relatively easy to observe changes in the number of users over time.

4. EFFECT OF VARYING INITIATION RATES

So far we have considered only a very simple model of initiation, specifically that initiation is 1,000,000 users per year times the current price raised to the price elasticity of initiation. In fact, initiation tends to vary over the course of a drug epidemic in ways that cannot be explained entirely by changes in prices for at least two broad reasons. First, drug users rarely start spontaneously; most are introduced to drug use by an existing drug user, often a friend or family member. This suggests that initiation should be an increasing function of the current number of users. Second, once use is widespread, the ill effects that manifest, particularly among heavier users, begin to give the drug a negative reputation that suppresses initiation. This has been described for past epidemics by Musto (1987), and has been noted for crack cocaine in the U.S. Hence, one can think of drug epidemics as having two phases. Initially the drug is seen as new and exciting, and initiation rates are relatively high. Later, initiation rates fall to include only those who would start using despite having witnessed the adverse effects of the drug on earlier cohorts. The next two subsections consider in turn how these variations in initiation affect the results.

4.1. Initiation as an Increasing Function of the Current Number of Users

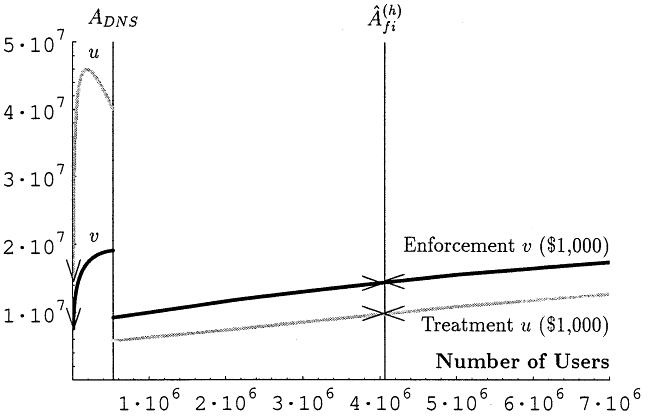
We wish to make initiation an increasing function of the current number of users. Presumably, this function should be concave because of saturation effects, so we replace the fixed initiation parameter k above with $k_2 A(t)^\alpha$. We choose $\alpha = 0.3$ and $k_2 = 4272$, because that gives the same rate of initiation at our base level of $A = 6{,}500{,}000$ users and a controlled steady state value ($\widehat{A}$) similar to that above for the constrained budget problem.

The analysis is analogous to that given above (for details, see Tragler et al. 1997). Modifying the initiation term does not change at all the qualitative behavior of the controlled system dynamics for the constrained budget problem. For the unconstrained budget problem, in contrast, the results are quite different, as can be seen from the phase portrait in the A-v-plane (Figure 6). In addition to the "high volume" saddle point equilibrium, there is a second "low volume" equilibrium that is an unstable focus, so the optimal policy is more complicated. For initial numbers of users above some critical level the solution is qualitatively just as before, a slow approach to the high volume saddle point equilibrium.

For smaller initial numbers of users ($A(0)$), it is not possible to jump onto the stable manifold that leads to the saddle point equilibrium. If we assume there is some lower limit, $\underline{A}$, on the number of users (e.g., $\underline{A} = 10{,}000$) below which enforcement and treatment cannot drive the problem (e.g., because these residual users cannot be detected), then the point $(\underline{A}, \underline{v})$ becomes another equilibrium, where $\underline{v}$ is given by the intersection of $A = \underline{A}$ and the isocline $\dot{A} = 0$. This steady state is approached along the trajectory which spirals out of the low volume equilibrium.

For low enough initial numbers of users it is only possible to jump on the stable manifold that approaches the lower limit equilibrium. For high enough values, it is clear one should approach the high volume equilibrium. For

Figure 7. Optimal treatment (gray) and enforcement (black) spending as a function of A; the two vertical lines indicate the DNS threshold and the high steady state value.



intermediate values, it is not immediately clear which is optimal, but there must be a so-called Dechert-Nishimura-Skiba (DNS) point (Skiba 1978, Dechert and Nishimura 1983, Feichtinger et al. 1997) that defines two basins of attraction according to whether the optimal policy is to effectively eradicate drug use (push it to the lower limit equilibrium) or to just moderate its approach to the high volume saddle point equilibrium, as above. For the base case parameter values, that point is $A_{\text{DNS}} = 529{,}117$ users.

Figure 7 shows the optimal amounts of treatment and enforcement spending as a function of the number of users. It illustrates several interesting points. First, if the initial number of users is to the left of the DNS point, treatment and enforcement spending are very high in absolute terms and, thus, truly enormous per user. If it is optimal to eradicate the drug epidemic, then apparently it is optimal to do so aggressively and quickly (cf. Baveja et al. 1997). This explains why in the constrained budget problem, at least with $G = \$1{,}600$, it continued to be optimal to allow the epidemic to grow to the saddle point equilibrium even with the modified initiation function. Second, to the right of the DNS point it is optimal to spend more on enforcement than on treatment, but the opposite is true to the left of the DNS point. Third, starting with $A(0) = 100{,}000$ users, the total social cost is less with the unconstrained budget than with the constrained budget ($613 billion vs. $838 billion) because by spending enormous amounts on control in the early years, it is possible to avoid getting stuck at the high volume equilibrium.

The larger the lower limit, $\underline{A}$, below which controls cannot drive the number of users, the smaller the DNS point. For example, doubling $\underline{A}$ to 20,000 roughly halves the DNS point (reduces it to 238,274). If the minimum number of users is interpreted as the number below which users are essentially invisible and immune to intervention, this has an interesting implication. Policy makers would like to push that lower limit down as far as possible. Doing so raises the DNS point and, thus, increases the time it takes an epidemic to reach the "point of no return," beyond which the best that policy can do is moderate expansion to the high volume equilibrium.

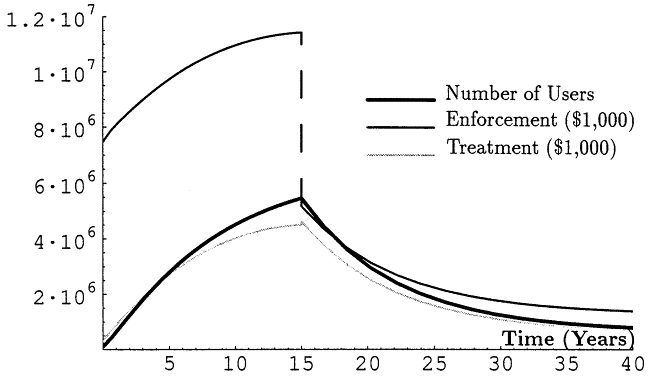
4.2. Managing the "End Game" after Initiation Has Fallen

So far we have primarily discussed what to do in the early and middle stages of an epidemic, when initiation is relatively high, and, unless one is to the left of the DNS point just discussed, the number of users is growing toward a high volume equilibrium or at that equilibrium. But demand for drugs among new users is cyclical, and there comes a time when initiation declines even if prices are relatively low and there are many existing users. For example, U.S. cocaine initiation rates dropped by 50% between 1988 and 1991 (Johnson et al. 1996). So we also analyzed a two-phase problem in which the initiation is at its base level for $T = 15$ years (starting with $A(0) = 100{,}000$ users) and then falls to 10% of its base value from then on (and in neither case is a function of the current number of users).

Analytically the main task is to glue the two problems together using a transversality condition which flows from the optimality requirements. The objective function in the first phase is evaluated from time 0 to time $T = 15$, not to infinity, but is augmented by a salvage value function, $e^{-rT}S(A(T))$, which is a function of the number of users at time T, i.e., $A(T)$. That salvage value is just the objective functional value for the original problem when the initial number of users is $A(T)$ and the initiation constant is cut by 90%. (For details see Tragler et al. 1997.) For the sake of brevity we discuss only the unconstrained budget problem results.

The time paths of the numbers of users (Figure 8) increase steadily but at a decreasing rate for the first T years. This growth is interrupted by the shift to a lower initiation rate and is followed by what looks like an exponential decay toward the low volume equilibrium. Recall

Figure 8. Plots over time of the number of users (A), enforcement spending (v), and treatment spending (u) for a two-phase problem in which the rate of initiation is cut by 90% after 15 years.



that the problem essentially scales in the initiation proportionality constant, so cutting k by 90% gives a new stable equilibrium that is only about one-tenth the size of the original equilibrium.

The time paths for the controls are similar with one big exception. Recall that the optimal level of enforcement is roughly linear in the number of users, i.e., $v^* \approx c_1 + c_2 A$. Cutting k by 90% sharply reduces the intercept, c_1, so at time T there is a sudden, discontinuous drop in the optimal level of enforcement. In contrast, although at time T, the optimal level of treatment spending stops increasing and starts to decrease, there is not a similar, abrupt change in the level.

Of course initiation rates do not actually drop by 90% overnight. The real world is not a two-phase problem but a multi-phase problem with steadily declining initiation constants in later stages of the epidemic. Nevertheless, this stylized example suggests that although it is optimal to use enforcement aggressively in the early, growth stages of an epidemic, later, once initiation has begun to ebb, it may be optimal to cut back enforcement, perhaps substantially.

5. DISCUSSION

The analysis above suggests some general observations. First, the traditional tendency to discuss the relative merits of different drug control interventions in static terms (e.g., claiming that treatment is better than enforcement or vice versa, without reference to the stage of the epidemic) is overly simplistic. Even a very simple model of drug use and drug control can, when viewed as an optimal control problem, yield optimal solutions that involve substantially varying the mix of interventions over time. However, the value of such flexibility depends on the level of control spending, and choosing the right level of control is more important than optimally varying the budget shares.

Second, inasmuch as our model is valid, in broad terms the policy recommendation is as follows. When policy makers become aware of the existence of a new drug problem, they must make a discrete choice as to whether they are going to try to eradicate the use of that drug or pursue a more humble policy of just moderating its growth toward a high volume equilibrium. Three conditions must be satisfied for eradication to be the optimal strategy. First, the drug epidemic must truly be an epidemic in the sense that new users are recruited by existing users, so that eradication pays dividends by short circuiting a positive feedback loop. Second, the problem must have been detected and reacted to early, before the number of users has grown beyond a critical threshold. Finally, policy makers must have the political capital necessary to direct very substantial resources toward the problem even though it is still relatively small, and the will to do so even though they might not be rewarded for averting an epidemic the public never saw. Indeed, political opponents might even accuse such a far-sighted leader of "crying wolf." If any of these three conditions is not satisfied, then an accommodation strategy is preferred.

If eradication is pursued, then it is optimal to move quickly and decisively, using truly massive levels of both enforcement and treatment to drive the number of users down as quickly as possible. If accommodation is pursued, then initially most energies should be directed toward enforcement. Over time, as the number of users grows, the level of enforcement should grow, but less than proportionately. Treatment spending, in contrast, should grow roughly proportionally with the number of users, so that over time treatment should receive a greater and greater share of the drug control budget.

Eventually most drugs develop a reputation for being dangerous and that deters initiation, even if the drugs remain relatively available and affordable. Once this stage happens, the number of users will begin to decline. Treatment spending should decline as the problem abates, but relatively smoothly. In contrast, once the drug has lost its attraction to most potential users, the value of enforcement's ability to keep prices high as a means of moderating initiation declines, so enforcement spending should be cut more aggressively in the end stage of the epidemic.

Our third observation is that there is abundant opportunity for further work. Extending the model to include more states (reflecting different types of drugs, different intensities of use for one drug, or the size of the susceptible population) and/or more controls (e.g., prevention) would be of interest. It would be valuable to link drug models to models of related social problems such as HIV/AIDS, property crime, and/or labor market outcomes. We modeled enforcement against suppliers as driving up equilibrium prices, not enforcement against users or enforcement that creates temporary conditions of physical scarcity. Likewise, we focused on treatment's ability to persuade people to quit and ignored treatment's incapacitative effects (reduced use during treatment even if followed by relapse) and its ability to reduce the harm per unit use (e.g., by reducing rates of HIV transmission). It would be important to determine whether our basic conclusions persist with richer models of the interventions.

Finally, it would be interesting to look in more detail at the role of treatment and enforcement in the very early stages of an epidemic. We found that if policy makers could intervene when the number of users was "small enough," then, if two other conditions were met, it would be optimal to try to eradicate drug use. But we say very little about how small is "small enough," and we are skeptical that any aggregate, market-oriented model could. Early distribution of drugs occurs primarily through social networks, not anonymous market transactions. It may be that an entirely different form of analysis is necessary for those crucial early years of a burgeoning epidemic.

ACKNOWLEDGMENTS

This research was partly financed by the Austrian Science Foundation (FWF) under contract No. P11711-SOZ

("Dynamic Law Enforcement"), the National Consortium on Violence Research, and by the U.S. National Science Foundation under Grant No. SBR-9357936. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors, and do not necessarily reflect the views of the National Science Foundation.

APPENDIX

The appendix can be found at the *Operations Research* Home Page in the Online Collection at ⟨http://or.pubs.informs.org⟩.

REFERENCES

Baveja, A., J. P. Caulkins, W. Liu, R. Batta, M. H. Karwan. 1997. When haste makes sense: cracking down on street markets for illicit drugs. *Socio-Economic Planning Sci.* **31** 293–306.

Becker, G. S., M. Grossman, K. M. Murphy. 1994. An empirical analysis of cigarette addiction. *Amer. Econom. Rev.* **84** 397–418.

Behrens, D. A., J. P. Caulkins, G. Tragler, G. Feichtinger. 1997. Controlling the U.S. cocaine epidemic: prevention from light vs. treatment of heavy use. Working Paper 214, Department of Operations Research and Systems Theory, Vienna University of Technology.

Caulkins, J. P. 1998. The cost-effectiveness of civil remedies: the case of drug control interventions. L. Green Mazerolle, and J. Roehl, eds. *Crime Prevention Stud.* **9** 219–237.

Caulkins, J. P., Reuter, P. 1997. Setting goals for drug policy: harm reduction or use reduction. *Addiction* **92** 1143–1150.

——. 1998. What price data tell us about drug markets. *J. Drug Issues* **28**(3) 593–612.

——, G. Crawford, P. Reuter. 1993. Simulation of adaptive response: a model of drug interdiction. *Math. Comput. Model.* **17** 37–52.

——, C. P. Rydell, W. L. Schwabe, J. Chiesa. 1997. *Mandatory Minimum Drug Sentences: Throwing Away the Key or the Taxpayers' Money?* RAND, Santa Monica, CA.

Crane, B. D., A. R. Rivolo, G. C. Comfort. 1997. An empirical examination of counterdrug program effectiveness. IDA Paper P-3219, Institute for Defense Analysis, Alexandria, VA.

Dechert, W. D., K. Nishimura. 1983. A complete characterization of optimal growth paths in an aggregated model with a non-concave production function. *J. Econom. Theory* **31**(2) 332–354.

Everingham, S. S., C. P. Rydell. 1994. *Modeling the Demand for Cocaine*, RAND, Santa Monica, CA.

Farrell, G., K. Mansur, M. Tullis. 1996. Cocaine and heroin in europe 1983–93: a cross-national comparison of trafficking and prices. *British J. Criminology* **36** 255–281.

Feichtinger, G., R. Hartl. 1986. *Optimale Kontrolle Ökonomischer Prozesse—Anwendungen des Maximumprinzips in den Wirtschaftswissenschaften*. deGruyter, Berlin.

——, W. Grienauer, G. Tragler. 1997. Optimal dynamic law enforcement. Working Paper 197, Department of Operations Research and Systems Theory, Vienna University of Technology.

Gerstein, D. R., R. A. Johnson, H. J. Harwood, D. Fontain, N. Suter, K. Malloy. 1994. *Evaluating Recovery Services: The California Drug and Alcohol Treatment Assessment*. National Opinion Research Center, Chicago, IL and Lewin-VHI, Fairfax, VA.

Hartl, R. F. 1987. A simple proof of the monotonicity of the state trajectories in autonomous control problems. *J. Econom. Theory* **41** 211–215.

Johnson, R. A., D. R. Gerstein, R. Ghadialy, W. Choy, J. Gfroerer. 1996. *Trends in the Incidence of Drug Use in the United States, 1919–1992*. U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration, Office of Applied Studies, Washington, DC.

Kleiman, M. A. R. 1993. Enforcement swamping: a positive-feedback mechanism in rates of illicit activity. *Math. Comput. Model.*, **17** 65–75.

——. 1988. Crackdowns: the effects of intensive enforcement on retail heroin dealing. M. R. Chaiken, ed. *Street-Level Drug Enforcement: Examining the Issues*. National Institute of Justice, Washington, DC.

Léonard, D., N. V. Long. 1992. *Optimal Control Theory and Static Optimization in Economics*, CUP, Cambridge.

Miller, T. R., M. A. Cohen, B. Wiersema. 1996. *Victim Costs and Consequences: A New Look*. National Institute of Justice, Washington, DC.

Moore, M. H. 1973. Achieving discrimination on the effective price of heroin. *Amer. Econom. Rev.* **63** 270–277.

Musto, D. F. 1987. *The American Disease*. Yale University Press, New Haven, CT.

Office of National Drug Control Policy. Various years. *The National Drug Control Strategy*. The White House, Washington, DC.

Reuter, P. 1983. *Disorganized Crime: The Economics of the Visible Hand*. MIT Press, Cambridge, MA.

——, M. A. R. Kleiman. 1986. Risks and prices: an economic analysis of drug enforcement. N. Morris and M. Tonry, eds. *Crime and Justice: A Review of Research*. University of Chicago Press, Chicago, IL.

Rydell, C. P., S. S. Everingham. 1994. *Controlling Cocaine. Supply Versus Demand Programs*. RAND, Santa Monica, CA.

Rydell, C. P., J. P. Caulkins, S. S. Everingham. 1996. Enforcement or treatment? modeling the relative efficacy of alternatives for controlling cocaine. *Oper. Res.* **44** 1–9.

Saffer, H., F. Chaloupka. 1995. The Demand for Illicit Drugs. Working Paper No. 5238, National Bureau of Economic Research, Cambridge, MA.

Skiba, A. K. 1978. Optional growth with a convex-concave production function. *Econometrica* **46** 527–539.

Tragler, G., J. P. Caulkins, G. Feichtinger. 1997. The impact of enforcement and treatment on illicit drug consumption. Working Paper 212, Department of Operations Research and Systems Theory, Vienna University of Technology, Vienna.

Varian, H. R. 1996. *Intermediate Microeconomics: A Modern Approach, 4th Edition*. W. W. Norton & Company, New York.