MODELLING DRUG MARKETS

National Institute of Justice
Project No. 89-IJ-CX-0001

Submitted by:

Program in Criminal Justice Policy and Management
John F. Kennedy School of Government
Harvard University

June 30, 1991

Principal Investigator: Mark A.R. Kleiman

Project Monitor: Dr. Bernard A. Gropper

MODELLING DRUG MARKETS: ABSTRACT

The economic analysis of drug dealing and drug policy continues to develop and deepen.

It has been argued that high-level enforcement efforts, particularly interdiction, will have only small effects on final retail prices and thus on consumption. That argument depends on the causal relationship between wholesale and retail drug prices. If much of the cost of retail drug dealing tends to be proportional, rather than merely additive, to the cost of drugs at wholesale, the effect of interdiction and other high-level enforcement efforts will be greater than previous analyses would suggest.

Since enforcement resources are scarce, the probability of enforcement action against any one transaction falls as the number of transactions rises. This external economy of scale reinforces upward and downward swings in drug volumes. Therefore, enforcement resources will be most effective if they are concentrated rather than dispersed and if the enforcement effort leads, rather than lagging, trends in physical drug volumes.

The search costs of drug purchase, and the neighborhood effects of drug dealing, are determined by retail market conditions. Enforcement efforts can be designed to make flagrant dealing impractical. An analysis of the factors of production in retail drug dealing can aid that design effort.

TABLE OF CONTENTS

MODELLING DRUG MARKETS - OVERVIEW

Mark A.R. Kleiman

MODELING DRUG MARKETS

Overview

Introduction: Toward a General Theory of Illicit Markets

Transactions involving certain goods and services are prohibited in every advanced society, though the list of prohibited items varies. The purchase and sale of foreign currencies and precious metals, sexual services, the labor of children and of certain classes of aliens, babies for adoption, contingent claims of various kinds (particularly those involving the outcome of sporting contests), human organs for transplant, loans not dischargeable by bankruptcy or carrying more than a legally specified rate of interest, methods of preventing or terminating pregnancy, and votes have all been banned at one time and place or another. The production and distribution of a long list of psychoactive substances for non-medical use are now prohibited almost universally by domestic law and international treaty.

The reasons for the prohibitions are almost as various as the prohibitions themselves. Usury laws are designed to protect borrowers (buyers). Child labor laws are designed to protect the underage workers (sellers). Prohibitions on bribing citizens and lawmakers are designed to protect the political process. Laws about selling sexual services, babies, and organs for transplant seem to be designed to insulate important social institutions from market forces. Currency laws try to maintain the control of the state over the economy.

A special category of prohibited transactions involves items and activities believed to be vices. A vice, in this usage, is a practice likely to have bad effects on the health, welfare, and character of those who engage in it, or who engage in it to excess. By altering the dispositions of their partakers, vices can damage both them and others, particularly if the vicious behavior carries within it a tendency to excess with respect either to frequency or intensity. Gambling, prostitution, and drug laws are all justified in part on vice-control grounds.

The mere existence of a prohibition suggests that, in its absence, there would be a market for the forbidden item: i.e., that some consumers would be willing to pay enough to induce some producers to provide it. Prohibition does not eliminate that potential market, but, like taxation, it inserts a "wedge" between the price (in money and non-money terms) the consumer pays and the price (net of enforcement risk) the producer receives.

If the prohibition is perfectly enforced, that wedge will be of effectively infinite size, and the would-be seller will find no buyer, the would-be buyer no seller, at any price. In the more general case, enforcement is imperfect, and the wedge of finite size, so that some buyers and sellers remain in the market and some transactions take place. Illicit-market economics is the study of those transactions and of the attempts to prevent them.

Prohibited transactions take place, in general, at higher money prices than would be true if the transactions were permitted, thus reducing the consumers' surplus (as evaluated by the consumers' tastes and opinions as of the moment of the transaction). In addition, the nature of illicit transactions will impose costs of other kinds on consumers: risk of apprehension and punishment, uncertainty and variability in product characteristics (e.g., adulterated or mislabeled drugs), lack of legal recourse for fraud or injury, and difficulty in obtaining market information (leading both to inability to comparison-shop and to the expenditure of time and other resources in the search for a seller). Putting to one side any psychic benefit the consumer might receive from violating the law, his gain in utility terms from each completed transaction is thus smaller than it would be if the market were a licit one.

Not all of this loss to the consumer is gain to the producer: thus the wedge. Only the higher money price benefits the producer, and some of this represents the additional cost of production and distribution in illicit markets, including not only the inconveniences incident to not being able to advertise, enter into legally enforceable contracts, and so on, but also the costs of either evading enforcement efforts or suffering the penalties of the law. Consequently, the volume of illicit production and consumption will in general be smaller than would be the case in the absence of prohibition.

If one accepts the judgment which underlies a vice prohibition in the first place -- that transactions in the commodity involved generate net losses when the real interests of producers, consumers, and third parties are all considered -- this reduction in the volume of transactions should be considered as a benefit of prohibition. The capacity of even weakly enforced prohibitions to produce this hypothetical benefit is demonstrated by the history of American alcohol prohibition. (Moore and Gerstein 1981; compare Morgan 1991).

The efficacy of any given prohibition in reducing the transactions it forbids depends on the interaction of three factors: the strength of demand and its elasticity to price and non-price conditions of availability; the individual dispositions and organizational capacities of actual and potential illicit producers; and the vigor and skill of the enforcement effort. Thus the most obvious purpose of prohibition enforcement is the support it lends to the purpose underlying the prohibition: in the case of drug law enforcement, the reduction in the volume of drug consumption and its related harms to users and others.

But that is not the only, or even necessarily the most important, purpose of vice enforcement, which can also serve symbolic, institutional-protective, and crime-control objectives. Symbolically, the arrest and punishment of vice producers and consumers reflects and reinforces public disapproval of the activity involved, and helps maintain the normative and deterrent force of other legal prohibitions. Since black-market dealing can threaten a wide variety of important institutions -- neighborhoods, workplaces, schools, and families -- the enforcement of vice laws can also help maintain the capacity of those institutions to perform their social functions. (Moore and Kleiman 1989) Insofar as black-market buyers and sellers are disproportionately represented among the perpetrators and victims of predatory crimes (all the varieties of theft, fraud, and assault), and the disorder incident to black-market dealing creates criminogenic conditions (Kelling and Wilson 1982) vice law enforcement can also serve to control crimes other than those defined by the vice laws themselves.

Thus the enforcement of vice laws ought to be designed both to effectuate the underlying purposes of those laws and to limit their unwanted side effects in the form of the social costs imposed by black market activity. These evaluative dimensions are sometimes complementary, sometimes competing. In particular, price-mediated reductions in vice consumption may be accompanied by increases in black-market side-effects. For example, higher heroin prices may lead to short-run increases in theft activity by heroin addicts as they struggle to maintain their habits. (Brown and Silverman 1974) Whether this is so will depend in large part on the price-elasticity of demand for the vice and the substitution behavior around it.

The presence of potentially competing objectives increases the value to decision-makers of conceptually accurate models, because models can help to reveal the value choices implicit in what appear to be merely technical decisions about tactics. Even in the absence of such tradeoffs, models of the vice markets can contribute to sound enforcement decision-making by illuminating the relationships between the choice of techniques and the application of resources on the one hand and likely results on the other. (Reuter and Kleiman 1986; Reuter, Crawford, and Cave 1988; Kleiman, Kulawik and Chayes 1990; BOTEC 1990; Kleiman and Smith 1990) They can also help illuminate resource-allocation choices between enforcement and other vice-control techniques: persuasion, treatment, and the control of problem users. (Moore 1990)

The more precise the models and the more available the data needed to implement them, the more uses to which models can be put. Given adequate models and data, it may be possible to predict the course of drug market developments, in terms of geography and drugs of abuse, to prevent a repetition of the surprise presented by the rise of the cocaine market in the late 1970s and early 1980s; in retrospect, some of those developments appear inevitable in light of data contemporaneously available. (Kleiman 1987) Such predictions can be useful both in planning responses to, e.g., the increasing workloads for criminal justice and social service agencies which followed the rise of cocaine and in planning interventions designed to prevent predicted unwelcome developments. Quantitative economic models of the drug markets could begin to answer the key question which confronts those who manage drug enforcement: "Can we expect this intervention to make a difference?" Moreover, market analysis is useful framework for evaluation, because it relates measures of desired outcomes (e.g., reduced transaction volumes) to measures of program outputs (e.g., arrests, seizures, and sentences). (Kleiman 1989; Kleiman et al 1988)

An economic analysis of drug-related behavior highlights certain questions which an epidemiological or criminal-conspiracy approach would not:

-- At what *prices* do various drugs trade, and how do those prices vary with transaction volumes? What determines those prices?

-- What is the total ***quantity*** purchased by final users? What is the relationship between quantity and price?

-- What are the ***revenues*** of drug-dealing organizations and the ***earnings*** of their principals and employees?

-- In addition to wages and entrepreneurial earnings, what are the ***costs*** of being in the drug-dealing business? What is the relationship between enforcement activity and the costs faced by drug dealers? (Reuter and Kleiman 1986)

Models based on economic analysis can help us interpret levels, trends, and sources of change in the drug markets: that is, they can be aids to ***description***. They can analyze the effects of current policies and programs on various aspects of the problem: as aids to ***evaluation***. They can help forecast the future, as aids to ***prediction***. Finally, they can inform ***policy planning*** by making a variety of contingent predictions of possible futures, using as assumptions alternative government actions. (Caulkins 1990; Kleiman Forthcoming)

The value of such models depends in part on the timely availability of data of sufficient detail and accuracy, and in part on the conceptual adequacy of the models. Since illicit markets both resemble and differ from licit ones, it is necessary to develop a body of illicit-market economics which takes those differences into account.

The economic analysis of drug markets can proceed in either or both of two directions: by observing actual market behavior, noting the peculiar features which distinguish drug markets from more conventional markets, and then seeking explanations for those peculiarities either in the illicit nature of the business or the special characteristics of drugs as consumer commodities; or by reasoning from general principles of economic analysis, combined with a few stylized facts about drug dealing (the transactions are illicit, buyers may diverge from ideal utility-maximizing consumers due to physical dependency, habituation, or binge cycles) to propositions about how the drug markets should be expected to behave.

Observational Characteristics of Drug Markets

1. **High Prices** Illicit drugs are far more expensive than comparable licit commodities, or than the same drugs sold in licit channels. Even if there were little or no enforcement, illicit goods would be more expensive than their licit counterparts because prohibition complicates doing business. Producers who cannot advertise, borrow from banks, sell securities, sign enforceable contracts, patent their inventions, or protect their trademarks will have to find clumsier, more expensive substitutes for these things which licit enterprises take for granted. In addition, the lack of advertising and trademarks makes it harder for customers to comparison-shop. In addition, illegality by itself tends to suppress consumption, independent of its effect on price, both because some consumers are reluctant to disobey the law and because illegal products are harder to find and less reliable as to quality and labeling than legal ones. Peter Reuter has referred to this complex of effects as "structural consequences of product illegality." To this is added the cost imposed on the illicit industry by the effort of enforcement agencies.

2. **High Money Incomes** Drug sellers earn higher money incomes than their skills would command as employees of licit, or most illicit, businesses. While their wealth has no doubt been exaggerated in popular and journalistic mythology, there is clearly good money to be made even at the lower end of the traffic. It appears

that low-level cocaine sellers in Washington, D.C., have average cash earnings of about $30 per hour, plus whatever drugs they use from their dealing stock, while the same individuals earn less than $10 per hour in their licit-market jobs. (Reuter, MacCoun, and Murphy 1990) Not all of this is pure gain, of course; against the cash earnings one must set substantial risks of imprisonment, injury in the course of business or business disputes, and the formation of unwanted and uncontrolled personal drug use habits.

3. **Dilution.** For some drugs, the actual contents of a package sold to the end-user is highly variable and only imperfectly known to the buyer. Changes in effective price are likely to take the form of, and be masked by, changes in purity.

Some of those mark-ups are obvious to the buyers. Others are partially concealed by the practice of "cutting," or diluting. The heroin content of a retail bag of "street heroin" through the 1970s and early 1980s averaged about 4% (even ignoring those instances where the customer was cheated by being sold a mixture not containing heroin at all). Heroin purity has been on the rise since 1986 as the wholesale price of heroin has fallen. In the late 1970s, retail cocaine tended to be about one-eighth cocaine, and the rest diluents; since then, purity on the street seems to have risen, and then fallen, with the fall and subsequent rise of wholesale prices. In the absence of regulations on packaging and labeling, the drug buyer has no assurance about the content of his purchase except the word of the seller; even after he has taken the drug, he may be unable to judge its quality with any precision Since price and quantity are directly observable and purity is not, dealers have an incentive for misrepresentation. In recent years, some heroin packagers have taken to using trademarks as a means of being able to build, and benefit from, a reputation for high quality. However, this tactic has had only partial success because the trademarks are not legally protected against competitors who substitute inferior goods for the genuine "Black Star" or "Murder One" brand.

The fact of cutting is well established, and some of the logic of it is clear. For example, if the dosages are small, dilution may make the drugs easier to handle. New sellers without established customer bases will not want to sell at greater purity than the average of existing sellers, since their claims of providing superior product are unlikely to be believed and they will thus be unable to charge premium prices. Rising wholesale prices increase short-term benefits of dilution to sellers. However, no student of the drug markets has ever produced anything like a convincing theory to explain how changes in market conditions cause changes in average purities over time or from market to market.

4. **Geographic Concentration** Drug transactions are highly concentrated geographically, almost certainly more concentrated than consumption, with a strong bias toward poor and socially disadvantaged neighborhoods. Why the markets are so concentrated, and why they are concentrated where they are, begs for explanation.

At the wholesale level, drug markets tend to be regional or even national in scope; for a period of several years, the largest cocaine distribution organization serving the Washington, D.C., market -- the network headed by Rayful Edmonds -- was receiving its cocaine shipments via Los Angeles, for no better reason than that Mr. Edmonds had a connection there. The costs of transportation, even with the precautions which enforcement makes necessary, are so small compared with the value of wholesale lots of drugs that large differences in prices among major markets are likely to attract arbitrageurs who will attempt to make a living on the differences and thus help shrink them. (This is least true for marijuana, which is bulky compared to its value.)

At the retail level, by contrast, relatively small differences in travel time, convenience, and perceived safety are likely to overwhelm price differences from the viewpoint of the user. The extreme example of this is provided by the heroin market in Manhattan. For more than a decade, the purity-adjusted price of heroin in Central Harlem was roughly twice its price in "Alphabet City" on the Lower East Side, less than an hour away by subway. Yet Central Harlem buyers, for many of whom heroin was the single largest item in their personal budgets, continued to buy close to home. Any attempt at arbitrage - buying in Alphabet City for resale in Harlem - would presumably have met with a violent reaction from the established uptown street-dealing organizations.

The desire of buyers to buy where they live would tend to spread retail drug dealing around, as it spreads convenience- store operations around. But working against this are the advantages of concentration. For the buyer, a street-corner with many dealers is a better place to seek drugs than a street-corner with a few, because the search time is likely to be less. For the seller, the picture is more complicated. On the one hand, competition is likely to drain away business. On the other, once a corner is known as a dealing corner, it will attract new buyers and have a better chance of maintaining its existing ones. Being a dealer on a busy corner is like being a physician in a group practice; going away for a week doesn't mean leaving your clientele hanging. But the great advantage of concentration, for buyers and sellers alike, comes from decreased enforcement risk. Sellers cluster for the same reasons fish shoal and birds flock: protection from natural enemies, in this case the police.

Since police routines tend to create a distribution of officers which is more uniform than the distribution of illicit activity, being the sole dealer on a corner is far riskier than being one of twenty. Buyers, too, insofar as they face enforcement risk, face much less of it in a crowd than they would alone.

5. **Search Behavior.** Buyers spend considerable time and effort in finding sellers, who may not be eager do business with all comers. The extent of consumer search varies considerably from buyer to buyer, and in some instances from day to day. Search behavior, and the distribution of search times, is both a phenomenon to be explained and a policy intermediate to be manipulated.

The fact that drug dealers do not advertise and have to be at least somewhat discreet about making their whereabouts known to prospective buyers means that the process of buying drugs often involves considerable searching on the part of the buyer, far more than is typical of licit markets. The search time for drugs -- which varies from buyer to buyer, from market to market and from time to time -- is a measure of what enforcement officials call "availability;" the longer the search time, the less available the drug is. But the term "availability" suggests a characteristic which is either present or absent, rather than a quantity which may be higher or lower. Some buyers who would be willing to pay the money price if drugs were easily available are unwilling to incur the effort and risk involved in looking for them (and the possibility of searching unsuccessfully). Search time thus acts as a second sort of price which users pay for their drugs. (Moore 1973, 1977)

Critics of enforcement efforts sometimes argue that such efforts are doomed to failure because "Anyone who really wants drugs can get them." Strictly speaking, this is true; at some price, after some amount of searching, any commodity that exists can (probably) be procured, and if one defines "really wanting" as being willing to go to any expense and inconvenience, then anyone who really wants any good can and will get it. The serious question for policy is how sensitive demand for drugs is to money price and search time, how effective various enforcement approaches can be in raising one or the other, and what can be done to reduce the willingness of buyers to spend time and money on drugs.

Increases in the prices of illicit drugs will do both good and harm, in proportions which vary with the responsiveness of demand to price. Higher prices will help suppress consumption, and presumably abuse. How much consumption falls as price rises (the price-elasticity of demand) depends on how attached various users are to the drug, what substitutes are available, and how large a role drug purchases play in their personal budgets. If the price-elasticity of demand is high -- if users purchase much less as prices rise -- then black-market price increases will generate substantial drug abuse control benefits. If the percentage change in quantity is greater than the percentage change in price (i.e., if the absolute value of the price-elasticity of demand is greater than one) total spending on drugs will fall as prices rise. Not only will this leave users including those who become former users, with more money to spend on food, clothing, shelter, and the support of their dependents, it will shrink the revenues of illicit-market dealers. Thus, where demand is elastic to price, higher prices do good all around.

For drugs in relatively inelastic demand, the results of price increases will be less happy. Consumption, and thus abuse, will shrink less than in the former case. Users will be somewhat less drugged then they would have been at lower prices, but poorer as their total expenditure on drugs rises. Those of them who finance their drug purchases by theft or illicit transactions (drug dealing or prostitution) may increase their criminal activity: "The drug squad makes work for the burglary squad." The revenues of drug dealers, and thus their capacity to pay for corruption and violence, will rise. If the frequency and seriousness of disputes among drug-dealers are roughly proportional to the total dollars involved, dealing-related bloodshed is likely to increase. (Since elasticity is higher in the long run than in the short run -- which means simply that habits change slowly -- the proportion of good to harm created by a price increase will rise over time. This suggests that steady, or steadily rising, prices are likely to create less damage than rapid fluctuations around an average.)

Increasing search times for retail drug purchases, like increasing prices, tend to reduce drug consumption. But high search times are free of the potential unwanted side-effects of high prices. Greater difficulty in buying drugs will lead to a smaller number of completed transactions, while leaving money prices unchanged. This will decrease both drug consumption and drug expenditures, leaving users with more money to spend on other goods and reducing their incentive to commit income-producing crimes. A heroin-using burglar or a crack-using prostitute facing a drug price increase can maintain his or her previous level of drug consumption by breaking into more houses or servicing more customers. But that strategy will not help in the face of an increase in search time, which makes it harder to turn dollars, including illicitly earned dollars, into drugs. Since smaller expenditures for the user means smaller revenues for the dealer, rising search times are unambiguously beneficial in controlling black-market corruption and violence as well.

6. **Frequent Purchases** Many buyers do not stockpile personal inventories, despite the presence of substantial discounts for volume purchase and the search costs involved in making many purchases. Since for frequent drug buyers the money spent on drugs represents a large fraction of their total personal budgets, this apparently irrational behavior requires some explanation.

This anomaly seems to be created, not by illegality or enforcement but by the nature of drug-using behavior. Given the very great gaps between prices at retail and prices only one step removed from the street (for example, a retailer pays $100 for between fifteen and twenty $10 bags of heroin) the large role of drug purchases in some buyers' personal budgets, and the great discomfort which some users face as a consequence of running out, we would expect consumers to attempt to buy in bulk and maintain personal inventories. This is widely true of marijuana users; but heavy users of heroin and crack rarely hold stockpiles. Partly this is simply a consequence of poverty and the risk of losing drugs to thieves or the police, but the greater difficulty comes from within. Many, if not most, heavy regular users of heroin and crack find it impossible to maintain personal "stashes" simply because they lack the self-command to save any drugs. The most common reason for ending a crack use session appears to be simply running out of crack, and strongly addicted heroin users sometimes resort to playing tricks on themselves to maintain enough for a "wake-up" shot. Even those users who are also retail dealers seem to have great difficulty in keeping themselves from dipping into inventory for personal use.

How much this relates to the nature of the drugs, and how much to the nature of the people who become heavy drug users under conditions of prohibition, it is impossible to tell. It does not seem to be characteristic either of marijuana or of the psychedelics, and it is less characteristic of powder cocaine than of crack. In any case, the result is that purchase units for heroin and crack are smaller, and transaction frequencies greater, than they would otherwise be. That magnifies the importance of relatively small increases in retail search time. If a marijuana smoker who buys a month's supply at a time finds that she has to search a few hours longer to find a seller, the pressure on her to reduce marijuana consumption is fairly small.

For the crack smoker who buys on as many days as he uses, or even several times in the same day, a change from a five-minute search to a forty-five-minute search may make a substantial difference.

Special Theoretical Considerations about Illicit Markets

Buyers and sellers in illicit markets face the risk of enforcement action and the inability to have recourse to law to enforce agreements and property rights (or even to complain of criminal victimization). The risk of enforcement action, and steps taken to reduce that risk, contribute to costs. Since any transaction exposes its participants to enforcement risks and the risk of violence or fraud by their transaction partners or others, transactions costs tend to be large, which helps to explain large markups from one distribution stage to the next.

Enforcement risks and the lack of legal recourse combined create high information costs for buyers and sellers. Not only is it difficult to find a potential transaction partner, but the cost of doing business with that person depend sharply on one's evaluation of his sincerity (i.e., not being an informant) and integrity (paying for what is delivered, delivering what is paid for). Buyers and sellers alike strongly prefer to deal with those they have dealt with before, even if the terms offered are apparently less favorable than could be obtained elsewhere. Thus the law of one price has far less explanatory power in the market for cocaine than in the market for cornmeal.

The inability to enter into contracts enforceable at law boosts agency costs (See Zeckhauser and Pratt 1991) Loan transactions are particularly problematic; the risk of the disruption of business by arrest and the seizure of assets increases the uncertainty about the borrower's ability to repay as the lack of legal compulsion raises the costs of ensuring that the borrower will repay even if he can.

Deprived of the coercive power of the state, illicit business people seek alternative means of resolving disputes. But without "a common power to keep them all in awe," it is difficult to make a dispute-resolution mechanism effective. Individuals and organizations with a reputation for coercive power and for the willingness to use it to enforce their decisions are therefore able to do business as dispute-resolvers. (Reuter 1983)

Behavioral Pharmacology and the Economics of Drugs

A theory of the drug markets must take seriously the behavioral pharmacology and sociology of drug use. Since these vary greatly from drug to drug, the generality of any theory of the drug markets will be limited, but some general remarks are possible.

Drug use is learned behavior for the individual and drug use practices are communicated within social groups. For most drugs other than the stimulants, the drug effect is perceived as pleasurable only after some experience in use, and the most likely source of that experience is a friend, relative, or acquaintance who is already a user of the same drug. (Kandel et al 1986; Huizinga and Elliott 1981) Thus higher drug consumption in one period and for one individual tends, other things equal, to increase consumption in future periods and for related individuals. These tendencies are common to many classes of consumer goods, but seem to be particularly marked for drugs, licit and illicit.

Thus a decrease in price today will, by increasing current consumption, tend to increase demand in the future. In addition, most mind-altering drugs create some degree of tolerance to their desired effects: i.e., after some period of use, the dose required to achieve the same level of drug effect will grow. The effects of tolerance are aggravated by the effects of dependence. Some users of almost all drugs, and many users of a few drugs (tobacco in the form of cigarettes, cocaine in smokable forms such as freebase or crack, heroin) enter into a state in which the cessation of drug-taking is physically or psychologically uncomfortable. (For example, heroin users typically suffer cramps, cocaine-smokers depression and an inability to experience pleasure.) The combination of tolerance and dependence -- addiction in its classical form -- can lead to the consumption of very large amounts of drugs.

The effects of addiction on aggregate consumption are complicated. Logically, someone addicted to an inexpensive drug with no close substitutes should have a very low price-elasticity of demand for that drug, and much discussion of the behavior of drug-users in general assumes that their drug-taking is determined entirely by their habit and not at all by price. Studies of cigarette-smoking confirm that consumption by current users changes much less in the face of tax increases than consumption by new users.

If, on the other hand, the drug is so expensive and the habit so great that it consumes a large proportion of the user's budget, then effective demand may become somewhat elastic simply due to the budget constraint. Moreover, the status of being addicted can be deliberately altered, at the cost of some discomfort, and there is some evidence that heroin users deliberately detoxify themselves when their habits get too expensive. (Kaplan 1983) Thus if changes in price influence the frequencies of initiation, quitting, and relapse, they may induce an significant elasticity in aggregate demand for even highly addictive substances even if they have only small effects on (non-quitting) current heavy users.

Moreover, the knowledge that a drug is addictive may act as a deterrent to trying it. The higher the expected price in the future, the greater that deterrent ought to be. A rational person considering whether or not to take an addictive drug should be more strongly influenced by a change in its price than he would be if the drug were not addictive, because the effect of the drug's price on his lifetime budget is greater.

The applicability of theories of "rational addiction" depends in part on the proportion of drug-takers who are in voluntary control of their drug-taking behavior. The fact that a drug may create addictive is not, as Becker and his colleagues have shown, adequate to establish that all use of it is irrational. (Becker and Murphy 1988; Becker, Grossman, and Murphy 1991) But there is powerful evidence that much drug purchase activity does not obey the principles of rational consumer choice.

Even in the absence of classical addiction, some proportion of the users of virtually any psychoactive will find that their drug-taking behavior is no longer under their fully voluntary control due to the effects of reinforcement. While they may not become "sick" if deprived of the drug (or after they have gone through a period of withdrawal and are physiologically "clean") they find that their behavior in the presence of the drug or in the presence of other "cues" to drug-taking is not the behavior they would choose for themselves in advance.

Persons who find themselves losing self-command with respect to drug-taking may seek to control their own behavior by removing themselves from drug-taking environments, by joining self-help groups, or by calling in professional help. It is hard to reconcile much of this behavior with the concept of a unitary rational actor maximizing satisfactions over a fixed preference ordering. (For a general discussion of the problems and strategies of managing one's own behavior, see Schelling 1984, chs. 3 and 4)

A major unresolved question about the behavioral pharmacology of drugs involves the relationships of substitution and complementary among them. Whether making one drug less available will, in the long run, increase or decrease the consumption of other drugs is crucial to making sound drug policy. The answer is probably different for different drug pairs. This is an issue on which theory is largely silent; empirical investigation is needed.

The Effects of Enforcement

In addition to the effects of illegality alone, black markets are shaped by the fact of law enforcement, which imposes costs and risks on suppliers (and sometimes on customers as well). Dealers risk losing their inventories, losing their non- drug assets, and losing their liberty by being arrested, convicted, and imprisoned. For those dealers who are employed by others, these risks form part of the background against which they must decide how high a wage to demand to start dealing or stay in the drug business rather than find a legitimate job, pursue some other illicit way of making money, or leave the labor market entirely. Those who are entrepreneurs must take enforcement risks into account in deciding whether to join, or remain in, the illicit industry. In either case, enforcement risks are part -- indeed, the dominant part -- of the cost structure of the illicit drug trade.

The magnitude of the enforcement risk depends on the size relationship between the enforcement effort and the illicit market. (Kleiman 1991) If market grows while the enforcement resources devoted to it do not, the result will be a smaller number of police and prosecutor work-hours and a smaller amount of prison space (or other punishment capacity) per transaction. If that reduced risk is passed on to the consumer in reduced prices, the result will be a further increase in volume, and thus a further decrease in risk, and so on. On a national level, the collapse of cocaine prices between 1979 and 1988 seems to have resulted in part from this effect. (Kleiman 1987) The same phenomenon is observable on a local level; one reason that drug markets tend to be concentrated rather than dispersed is that there is safety in numbers.

The tendency of growing drug markets to outstrip enforcement capacity creates what economists call "inter-firm economies of scale;" the larger the market grows, the less its products cost. That is one reason that drug consumption grows quickly during the early stages of a "drug epidemic" and fades quickly after its peak; the markets tend to reinforce demand swings with price swings. In terms of enforcement policy, this effect puts a premium on detecting markets at the early stages of their growth and moving enforcement resources quickly to meet new threats as they arise: the strategy which Churchill, in another context, called "strangling the baby in its crib."

Illegality and enforcement create markets where "transaction costs" -- the expenses in money and time associated with buying and selling -- are large compared to the underlying value of the goods involved. In the case of heroin and cocaine, more than 90% of the final retail price is added after the drugs land in the United States. That mark-up represents the labor and risk of domestic distributors and retail dealers.

General Economic Principles

Economists have deduced a small number of theorems of great generality regarding markets in which there are many buyers and many sellers and in which information flows freely. How, and to what extent, these theorems apply to illicit markets is a central question.

The first of these principles is sometimes referred to as "the law of one price." It holds that identical commodities will trade at (almost) identical prices; to be precise, that the difference in prices between two markets for the same commodity will not exceed the cost of moving the goods from one market to the other.

The justification for this principle in ordinary markets is straightforward. If sellers prefer to earn as much as they can, and buyers to pay as little as they can, then price differentials will generate behavior which will cause them to shrink. Goods will be carried from markets in which they are cheap to markets in which they are dear, while buyers will move in the other direction. Thus in the cheap market the supply will shrink while the demand expands (tending to increase the price there) while in the dear market the supply will grow while the demand shrinks (driving the price down). The greatest sustainable price difference, then, will be the cost of the transaction (understood to include the costs of learning about the price difference and finding persons from whom to buy and sell). It is not necessary, for this "law" to hold, that all participants be perfectly rational, or even that they all know of the existence of multiple markets, because of the possibility of "arbitrage" behavior: trading on the price differential by purchasing in the cheap market and selling in the dear one.

Another general principle of market behavior is the "zero pure profit" theorem. Competition prevents price gouging by driving prices down to the level at which no resource-owner can make more money by moving his resources into the industry than he could be applying the same resources to another opportunity. While firms in an industry have a common interest in higher prices, an agreement among them to do so will fail unless they have a way to enforce the agreement among themselves and have a way of keeping new players out of the game.

The zero-pure-profit theorem implies that, over the long run, the total paid by the consumers of a good is equal to the total cost of producing that good. In the absence of special circumstances such as barriers to entry, or economies of scale (that is, situations in which the average cost of producing a unit of some good falls as the total quantity produced by a single firm rises), the total revenue received by an industry will tend to equal its total costs (raw materials, wages, payments for special skills and resources, and the cost of capital, including both the rate of interest on risk-free loans and a premium reflecting the risks to capital providers).

As with the law of one price, the zero long-run pure profit theorem is supported by the observation that any deviation from it would allow someone to benefit himself, and that in the process of doing so he would tend to reduce the deviation. If prices in a competitive industry were such that long-run pure profits were available to be made, the factors of production (such as labor and capital) would tend to flow into that industry to share in the supra-normal returns, thus expanding supply and tending to reduce price. If prices were so low that costs were not being covered -- if the long-run pure profits were negative -- productive resources would tend to flow out of that industry into other industries where they were better compensated, thus contracting supply and increasing price.

This implies that, in a competitive industry, long-run price equals long-run cost, and whatever increases cost for the producer eventually increases price to the consumer.

The actual process by which market participants adjust themselves to external changes is a complicated one, which goes under the names "dynamics" or "disequilibrium behavior." But economists have gotten considerable mileage out of reasoning about the way markets will look once the dynamics have settled out and equilibrium is restored. The law of one price and the zero long-run pure profit theorem are both aspects of this idealized world, which goes under the name of "comparative statics" because its method is to compare the resting states of markets under alternative assumptions about external conditions.

Comparative Statics and the Evaluation of Drug Law Enforcement Programs

If the drug markets obeyed the rule that long-run pure profits must be zero, the market's total revenues would equal its total costs. In particular, the direct and indirect costs imposed by law enforcement would all be passed through to consumers. That would allow a fairly straightforward evaluation of at least that part of the law enforcement effort designed to decrease consumption through increasing prices: the important evaluative dimension would be (marginal) dollars of cost imposed per (marginal) enforcement dollar spent. This is the approach taken by Reuter and Kleiman. (1986)

Monetizing Enforcement-Generated Costs

The chief difficulty in actually performing such an evaluation would be in assigning dollar values to the four major components of cost imposition: drugs seized, assets seized, prison time, and the costs of avoiding these (precautions, bribes, lawyers' fees, etc.). Evaluating each of these poses conceptual as well as practical problems (Kleiman 1989, pp. 77- 82).

Cash, which accounts for the bulk of non-drug asset seizures, is easy to evaluate. Drugs and other non-monetary assets are trickier. As a theoretical guideline, seizures ought to be evaluated at the traffickers's costs of replacing them. This applies alike to seizures of drugs and of non-drug assets. In the case of hard non-drug assets such as boats, traffickers' replacement cost may be greater than the sum the government realizes from an auction sale. In the case of drugs, replacement cost will in general be far less that the final retail or "street" value often cited by the press; the further "up the chain" the transaction occurs, the smaller the per-unit replacement value. (An attempt to compare various enforcement efforts in terms of drug volumes seized per dollar spent (Wharton Econometrics 1987) has been vigorously criticized on this point. (Reuter and Cave 1988)

Prison time is even more difficult to evaluate. Conceptually, the question to be asked is, "How does the overall distribution of prison terms among drug dealers influence the costs of drug distribution?" It does so, presumably, through the behavior of those threatened with prison; employees will demand higher wages, and entrepreneurs higher proprietary earnings, as the threat of imprisonment rises. This phenomenon is parallel to the influence of industrial-accident rates on wages. (Viscusi, 1983) An important difference is that industrial accidents happen suddenly, while the process of investigation, arrest, trial, sentencing, imprisonment, and eventual release is extended in time. Therefore, while the rate of fatal industrial accidents in the steel industry this year can be determined this year, the number of years of imprisonment served for cocaine deals this year will not be determined for more than a decade into the future. This leaves the question of how market participants judge their imprisonment risks, and how different governmental actions (passing laws, making arrests, hiring agents, building prisons) influence that judgment. (The parallel here would be to the "rational expectations" literature and debate.)

Even if we assume that the number of years eventually to be served is somehow correctly judged by current and potential market participants, there remains the question of what dollar values they put on various risks of various sentences. Are there fixed costs of entering prison, independent of the length of time served? After that threshold, is a sentence twice as long more or less than twice as unpleasant? What is a year in prison worth, compared to a year in reduced life expectancy? (See Howard 1979) Most of all, how do these values vary from participant to participant, and specifically with the participant's income? This remains a challenging area for several different kinds of research: rational-actor-hypothesis calculations of optimal behavior ("vicarious problem-solving"); survey research among drug-market participants both to ask about their perceptions of risk and (using contingent-valuation techniques) their evaluation of that risk; and behavioral studies of the effects of changes in the severity and probability of sanctions on dealers' wages and earnings.

The Comparative-Statics Challenge To Interdiction

Despite the difficulties outlined above, it is not very difficult to assign roughly accurate monetary values to a variety of enforcement activities. In particular, data from undercover transactions and negotiations provide a reasonably clear picture of the replacement cost of seized drugs. The resulting calculations of costs imposed on the illicit industry tend to place a much smaller value on bulk seizures of drugs in smuggling vessels or container freight than is assigned to that activity in the popular imagination, or reflected in the share that interdiction programs take in federal drug enforcement spending. (Reuter 1988; Reuter, Crawford, and Cave 1988)

A shipment of cocaine seized in transit from its source country to the United States, or at the U.S. port of entry, is certainly worth less to the traffickers who own it than they would be able to sell it for once it had been safely landed, but more than it could be replaced for in its source country, since time and money have been expended on its transportation north. With per-kilogram prices in U.S. cocaine-trafficking centers hovering between $20,000 and $30,000, and export prices at about one-tenth of those figures, assigning a replacement cost of $10,000 per kilogram to a cocaine seizure would be generous.

Total interdiction seizures of cocaine in the years 1988-1990 averaged less than 85 metric tons (thousands of kilograms) (Rhodes and McDonald 1991), which is a respectable fraction of the roughly 200 metric tons estimated to be consumed per year in the United States; between a quarter and a third of all drugs shipped are intercepted at or before the border. If one had the view that bulk drugs at export were scarce, and that seized shipments were replaced, the interdiction effort would appear to be substantially reducing the cocaine problem in the United States over what it would otherwise be.

But if interdiction is evaluated as a means of imposing costs on the illicit industry, it seems far less impressive. Even at $10,000 per kilogram, and even if 100 tons per year were consistently taken, the total impact on the illicit industry would be $1 billion, out of estimated retail revenues of $17.5 billion. (Rhodes and McDonald 1991) Thus the comparative-statics analysis tends to be quite discouraging about the proposition that interdiction efforts can substantially raise the prices of imported drugs.

If in fact the effect of interdiction on price is small, and if (as seems beyond dispute) supplies of bulk drugs in source countries are no constraint on the traffic, then the effect of interdiction on consumption will be correspondingly small. This in turn implies that the total amount of cocaine shipped from source countries to the U.S. will tend to rise along with the success of the interdiction effort, as source-country suppliers sell enough to meet the (largely unaffected) final retail demand plus the "demand" represented by seizure. Since it is the export demand which determines the political threat drug trafficking poses to source countries and the environmental damage done there by cocaine growing and refining, an increase in export demand is disadvantageous to source countries. Thus if the comparative-statics or "risks and prices" view of the drug markets is substantially correct, interdiction has small benefits for the United States and imposes substantial costs on source countries.

Challenges to The Comparative-Statics Model

The comparative-statics model is not the only possible account of how drug quantities and prices are determined.

There are at least two competing accounts of the determination of physical volume, relying on variables other than price: the physical supply of drugs and the throughput capacity of drug-dealing organizations. These models lead to radically different prescriptions for enforcement.

Even if one accepts that drug markets are characterized by conventional supply and demand relationships, there are at least two objections to a straightforward "risks and prices" model: one based on the theory that retail prices may increase as a multiple of wholesale prices, rather than by adding relatively fixed costs of retail distribution to changing prices of bulk drugs, the other denying the applicability of the zero-pure- profit theorem to the drug markets due to the existence of organizational quasi-rents based on heterogeneity in cost structures.

Other Models of Quantity Determination

The Physical-Flow Model

One simple model of quantity determination is that the volume of drugs consumed depends only on the quantity produced in source countries and the quantity seized by enforcement. The implicit equation is "production (or attempted import) minus removals equals consumption." On this theory, enforcement does its work by removing drugs from the supply; whatever is seized would otherwise have been consumed.

The fallacy of this line of reasoning is itself almost self- evident: it assumes production (or attempted import) is determined independently of removals. But if the results of removals is unmet demand, what is to keep either current dealers or new entrants into the market from satisfying that unmet demand by bringing more drugs to market?

The drug removal theory would make sense only if the supply curve for the drug market was vertical or close to vertical. This would mean that the capacity of the world to produce raw drugs (or at least to produce drugs acceptable to consumers at something like current prices) was somehow fundamentally limited. But there is no reason to believe this. Official estimates of the drug production at the farm level outstrip the estimates of drugs finally consumed. (NNICC 1989; Rhodes and McDonald 1991) Moreover, if bulk drugs were in short supply one would expect the farmgate price to be a more significant proportion of prices further down the chain, that appears to be the case.

Organizational Capacity Models

Even if drugs cannot be in long-term short supply, there might be limits on the ability of drug dealers to acquire them where they are grown, arrange and finance their shipment to the United States, and distribute them to the final consumer down a chain of middlemen, none of them able to be sued and any of them capable of betraying his supplier to the authorities.

The difficulty and risk of finding trustworthy transaction partners in an illicit business might act as a "barrier to entry" keeping new drug dealing firms out of the market and as a constraint on the growth of existing firms. If this were the case, breaking up existing patterns of dealing among those who have learned to trust one another might reduce the overall capacity of the drug supply system. (Moore 1979)

Enforcement can disrupt such patterns and retard the development of new ones in three ways: by putting important participants in prison, thus preventing them from dealing during their stay behind bars; by causing participants to distrust each other either in particular (an individual under intensive investigation or indictment has strong incentives to betray his colleagues in hope of leniency) or in general (the more undercover agents there are pretending to be drug dealers, the riskier it is to buy or sell to a proported dealer); by confiscating the drugs and money which are to be "working capital" of ongoing drug-dealing organizations : if the existence for such an organization is what creates trust , and if that organization can be put out of business by the loss of working capital, taking assets may effectively destroy trust.

The through-put capacity theory implies that any particular moment the supply of drugs is limited, and increased enforcement reduces this supply. Enforcement can thus restrict supply and raise prices if it can succeed in disabling one or more large supply organizations.

This account is internally coherent, and it may explain some of the short-term behavior of the drug markets in response to unexpected external shocks, but theory and evidence suggest that it is unlikely to capture most of the truth about the drug markets most of the time.

If throughput could not adapt to changes in enforcement or demand, then a string of enforcement successes in one area should leave a local supply shortfall. Consumers would either face sharply increased prices (as dealers ration limited supplies by price) or be unable to buy at all (if dealers use nonprice rationing). Yet enforcement estimates of drug prices and their regional distribution, are rather stable over time.

Other Models of Price Determination

Multiplicative Pass-through

The observation that drug prices at export and import are a small proportion of final retail prices does not immediately entail the conclusion that *changes* in prices at those level have only a small *effect* on retail prices. If, for example, the price of heroin at retail were ten times the price at the kilogram level, and if that relationship continued to hold true regardless of changes in kilogram prices, then a dollar added to the cost of doing business at the kilogram level would cause a ten-dollar increase in the aggregate retail-level revenues of the heroin industry.

Caulkins (1990) points out that the ratio of retail to wholesale cocaine prices remained remarkably constant during the 1980s even as the prices themselves fell dramatically. This seems to support a model where wholesale price changes are passed through multiplicatively, rather than additively, at retail. However, another interpretation, more consistent with the comparative-statics or "Risks and Prices" model, is also available: that both wholesale and retail prices fell due to the same cause. Until the end of the decade, the physical volume of the cocaine traffic seems to have grown considerably more quickly than did the enforcement resources directed at it; this was true at all levels of the traffic. A falling ratio of enforcement effort to physical volume would be expected to lead to falling prices. (Kleiman 1991) Thus retail prices might have fallen along with wholesale prices without falling because wholesale prices were falling.

The challenge to those who propose a multiplicative relationship between wholesale and retail prices is to explain why. Some of the costs of retail dealing clearly are proportional to the value of the commodity: the cost of working capital to finance transactions, for example, and the value of drugs lost by seizure further down the distribution chain. But Reuter and Kleiman (1986 p. 305) calculate these obviously multiplicative costs as generating an additional price impact at retail of no more than several percent. Most of the costs of retail drug dealing seems to be the wages of the dealers; why should these wages increase with the price of the wholesale commodity?

Caulkins (1990) offers two answers to this challenge. First, a major cost of drug dealing is protecting the drugs from theft. Not only are more expensive drugs more costly to lose, they are also more worth stealing; security costs will therefore rise more than additively with changes in wholesale prices.

Second and probably more important, there are reasons for the labor costs of retail drug dealing to rise and fall with the value of the goods. Many retail dealers are paid in drugs; thus the cost of employing them rises automatically with wholesale drug prices. (The same is true of drugs used out of inventory by the operator of the retail dealing organization, or given by him to his friends.)

In addition, many employees of retail dealing organizations -- consignment sales personnel in particular, but also guards, couriers, and those who dilute and package the drugs -- have frequent opportunities to abscond with drug inventories. As with external theft, higher wholesale prices raise both the cost to the organization from employee theft and the incentives employees have to steal.

In the absence of recourse to law, employers in the drug trade have two primary ways of dealing with this problem. One is to threaten violent retaliation, and incidents of such violence are far from rare. But such violence is exceedingly expensive: it risks counter-violence and severe criminal penalties, it involves paying "enforcers" and then living with the threat they pose to life and property, and it creates a disincentive for employees which may have to be compensated for by higher wages. The second (possibly complementary) approach to ensuring employee honesty is to pay sufficiently supra-market wages that the risk of dismissal creates a substantial disincentive to dishonesty.

In effect, each employee will repeatedly face a choice between absconding with some drugs, gaining their value at once and with high probability, or continuing to work for the organization and collecting a stream of uncertain benefits -- wages in excess of the next best alternative -- over time. The employer's problem is to set wages high enough that the expected present value of future benefits to workers who do not abscond is, in the vast majority of cases, greater than the value of the drugs they can abscond with. That critical-value wage rises directly with the wholesale price of the drug. (Caulkins 1990)

The payment of what appear to be supra-market wages to induce employee honesty is not unknown even in licit businesses where dishonesty is expensive and hard to detect. Bartenders, for example, seem to be paid more than the difficulty of learning that trade can easily explain. Reuter (1990) gives evidence that retail cocaine dealers in Washington, D.C., earn on average the equivalent of $30 per hour in cash, in addition to payment in drugs. While less spectacular than the mythical wealth of drug dealers, this is still a higher figure than even the job's substantial enforcement risks would justify, and in fact Reuter reports that many dealers in his sample appeared to be "under- employed," working as dealers part-time for want of full-time employment opportunity in the drug trade. The failure of that labor market to "clear" at a lower wage suggests a departure from a perfectly competitive labor market, and thus lends indirect support to a Caulkins-style analysis.

Nothing less than the value of a large part of the federal government's drug law enforcement effort is at stake in this debate over whether the relationship of retail price to wholesale price is almost additive (with a small multiplicative term) or almost multiplicative (with a small additive term). To date, the question remains open, with arguments and evidence on both sides.

There is room for more research of several kinds: additional theoretical development, case studies of particular dealing organizations, and the examination of more recent time series of wholesale and retail prices, particularly during the bulge in wholesale cocaine prices which temporarily reversed a decade-long trend in mid-to-late 1990. More systematic collection of retail price and purity data than is provided by the DEA STRIDE system may be needed, and could be arranged for relatively little expense. (Reuter and Haaga 1991)

Dealer Heterogeneity, Learning, and Quasi-Rents

The theorem of zero long-run pure profit will not hold if some market participants own scarce factors of production which allow them to deliver commodities at lower cost than others. The application of this observation to more and less productive agricultural property led to the classical theory of the rent of land, and therefore economists call supra-market returns derived from resource ownership "rents." The analysis of rent has also been applied to the ownership of other natural resources (e.g., high-quality mineral deposits). In each case, the resource owner can receive as a rent the difference between the cost of producing a good using the highest-cost resource for which there is still demand (the marginal resource) and the (lower) cost of producing it using the resource which he commands.

This exception is easily accommodated without sacrificing the generalization that long-run pure profits in a competitive market must be zero by treating the rent as a separate factor payment, even if the rent-producing property is owned by the entrepreneur. The entrepreneur receives zero pure profit in the sense that his rent-producing property could have been sold to a third party, and his profit over and above the rent he could have received will be zero.

The analysis is only slightly more complicated when the rent-producing asset itself is produced rather than being found. A piece of physical productive capital, once produced, commands a rent; i.e., its revenue is greater than its operating cost. But in the long run, on average, the rental value of the equipment cannot exceed the costs of its production, or resources would crowd into the market in producing capital goods. Thus the rental on capital equipment is not a pure economic rent unless the analysis is confined to the present time; economists call such receipts "quasi-rents." The superior earnings ability of workers who have learned skills ("acquired human capital") can also be treated as quasi-rents. So can the royalties on intellectual property such as copyrights or patents.

In each case, even if the physical or human capital is possessed by the entrepreneur, its rental value can be separated out from the entrepreneur's earnings for the purpose of showing that no pure profit is derived; he could have hired out himself or his capital equipment to another entrepreneur, and derives no profit above his total costs including the forgone wages or lease payments.

But some rent-producing resources are inseparable from the firm which employs them. The web of business connections which accountants call "goodwill" may allow a firm to produce at lower cost than its marginal competitors, and thus to collect a pure profit over and above the cost of the labor and capital it employs. Its previous operations place it in a position to earn quasi-rents, even without possessing a specific vendable asset separate from the firm itself.

In the abstract and in theory, the accumulation of goodwill, like the accumulation of physical capital, can be regarded as part of the earnings of the firm involved. Just as a firm which runs at break-even on an economic basis may experience a negative cash flow in a period during which it acquires a piece of physical capital, one can imagine firms choosing to run deficits to build up goodwill. But the difficulties in measuring, financing, or selling goodwill make it difficult to deal with empirically.

Goodwill -- what Reuter has called "relational capital" -- is likely to be particularly significant in illicit industries, including drug dealing. Much of the enforcement risk in illicit transactions derives from the probability than the person on the other side of the transaction is an agent or informant, or may become an informant in the future. Since any transaction with a new buyer, seller, or employee adds to the number of persons whose testimony can put one in prison, while transactions with previous transaction partners do not, new and growing firms will face higher costs than old and stable ones.

While the technical reasons are different, the accumulation of relational capital ought to have some of the same effects on industry structure and pricing behavior as the decline of unit costs with aggregate historical production observed in high- technology manufacturing: the famous "learning curve." But while an airframe manufacturer can continue to reduce costs by expanding production, a cocaine dealer who tries to expand in volume faces new risks, because larger volume almost inevitably means doing business with new partners. A firm which grows too quickly risks losing its established profitable position.

If this is so, drug firms, particularly firms in growing drug industries, will tend to be heterogeneous in their cost structures, and established firms will be able to collect quasi-rents on their relational capital up to a volume level not too far removed from their historical experience. (One would therefore expect market participants to accept sub-market returns for early transactions in the expectation of reaping rewards later. This will not mean running losses on a cash basis, but accepting short-run entrepreneurial returns insufficient to compensate for the short-run risks of arrest, imprisonment, and the loss of assets.) The prices of drugs will reflect the costs of the newest, and therefore highest-cost, participants.

In this circumstance, enforcement directed at established firms may result in reduced quasi-rents rather than higher prices. On the other hand, the failure to destroy a sufficient amount of low-cost drug distribution capacity may lead to a price collapse, as existing firms whose aggregate capacity exceeds demand at the new-entrant price cut their margins due to competitive pressures from one another. (See Kleiman 1989, ch. 4)

In the very long run, if drug market participants are able to predict levels of enforcement activities correctly, the expectation of reduced future quasi-rents will boost current drug prices, as the willingness of current firms to accept sub-market returns in the short term declines. But the specification of the way in which enforcement today influences current and future expectations about enforcement is virtually impossible, and will greatly complicate the problem of using data to derive parameters such as the cost impact of an additional prison-year served by high-level dealers.

Industry-Wide Economies of Scale and Positive Feedback

The theory of multiplicative pass-through tends to emphasize the value of high-level enforcement in general and of interdiction in particular, compared to the value those activities have in a pure comparative-statics model. The theory of firm heterogeneity and the learning firm tends to complicate the picture, with policy recommendations highly sensitive to market conditions. But both accept the notion that the costs faced by a firm depend in the firm's behavior and the resources and tactics employed by enforcement agencies. This many not in fact be the case.

If enforcement resources are finite in the short run, the enforcement risk faced in the course of any one transaction will fall as the number of transactions rises, simply because the police cannot arrest everyone at once. As a result, an increase the volume of any one dealing organization will reduce the per- unit costs of all competing organizations: an effect economists call "industry-wide external economies of scale."

The dynamic thus created forces drug market economics beyond comparative statics. Industry-wide economies of scale will tend to reinforce upward and downward movements in drug volumes. Larger volumes today will lead to lower costs for dealing organizations and thus lower prices. Lower prices will induce increased consumption. Increased consumption today will lead to increased demand in the future, as a result of they physiology of drug use (tolerance and dependence), the individual psychology of drug use (reinforcement), and the epidemiology of drug use (current drug users, particularly recent drug users, help initiate new users). Increased demand will lead to still higher volumes, and so on. The same mechanisms will be in force on the way down. (Kleiman 1991)

This analysis has little to say about the relative value of interdiction and other high-level enforcement activity measured against competing uses of the same resources. But it raises questions about the allocation of enforcement resource across drugs, suggesting that it would be desirable for enforcement resources to lead, rather than lagging, trends in physical volume. From this perspective, the massive resources applied to enforcing the cocaine laws in the latter half of the 1980s would have been far more effective had they arrived earlier, when marijuana and heroin continued to dominate enforcement attention as cocaine prices fell and volume swelled. Attention to positive-feedback effects would tend to support arguments for the rapid movement of enforcement resources into the heroin market, which appears to be in a resurgence.

The Persistence of Markets for Illicit Goods

Drug markets have both inertia of rest and inertia of motion. Both current levels and current trends tend to perpetuate themselves. This is so for at least six different reasons.

First, current demand is a function of past consumption due to reinforcement, tolerance, and dependence, and also to the fact that the enjoyment of drug-taking is largely learned behavior.

Second, current supply is a function of past sales. Individual and organizations build the skills and inclinations to deal in drugs. Cocaine money may build as much tolerance and dependence as cocaine itself.

Third, as drug users learn to use drugs and drug sellers learn to sell them, they also learn to do business with one another. The relational capital among them -- their built experience of dealing with one another and the existence of a community within which reputations can be earned and lost -- is a common-property resource for the entire industry, and it tends to grow with time and volume.

Fourth, some of the important transactions costs of the drug business -- labor time and inventory holding time for dealers, search time for buyers -- tend to fall as the number of transactions rises and the experience of buyers and sellers grows.

The more buyers and sellers, the shorter the search required to find one or the other. The reduction in transactions costs reduces both components of users' "effective price" (money and search time) and thus increases volume. Over time, dealers learn whom to sell to, buyers whom to buy from.

Fifth, as enforcement resources are spread over more and more transactions, the enforcement risk of consummating any one transaction falls.

Sixth, since recently initiated drug users are most likely to be excited about the perceived benefits of drug use and least likely to show its ill effects, they are the most potent sources of positive word-of-mouth information and the most likely to initiate their friends. A growing drug market is likely to have many recently initiated users, and they will help it continue to grow, while a shrinking market will have few.

Against these six factors creating persistence in the short run are two which create counter-pressures over the longer term. As the number of persons who have been using a given drug for a long time grows, so does the number who are feeling and showing its negative effects. Not only are they no longer highly motivated or persuasive recruiters of new users, they actively discourage new use by their bad example, whether observed individually or communicated through the mass media. Thus a few years after a beginning of boom market in a particular drug, its reputation tends to sour. This reduces the number of new users, thus further reducing the number of proselytizers.

Growing drug markets also attract additional enforcement resources, due both to the purely professional desire of drug agents to make big cases and to pressure from the public to "do something." Enforcement resources are likely to continue to grow even after the market itself has peaked. This tends to accelerate the downward trend in volumes by further increasing the enforcement-to-transaction ratio.

The combination of short-term positive feedback and long- term negative feedback tends to create a multi-year cycle of boom and bust. There is evidence that the cocaine market may have reached its peak for this cycle in the past year or two. (For the historical analysis, see Musto 1987, 1991; for evidence on cocaine, see National Household Survey 1990)

Volume Determination and the Importance of Retail Conditions

Whatever the correct model of wholesale price determination, it is less than the entire story of the drug markets. Conditions of retail sale also play an important role in determining drug volumes. It is also retail drug dealing which creates the disruptive conditions which make open drug markets so devastating to the neighborhoods in which they occur, including a large proportion of the violence associated with the drug trade.

For some purposes, where and how drug dealing takes place may be as important as, or even more important than, the number of users or the quantity consumed. The attempt to influence retail conditions engages a different set of enforcement tactics than the attempt to influence wholesale prices. (Moore 1973; Moore 1977; Kleiman and Smith 1990; BOTEC Analysis Corporation 1990; Burns and Conner 1991). In some instances, concentrated retail-level enforcement activity has substantially improved the local quality of life where it has taken place. (Kleiman 1988; Zimmer 1987)

The first to apply economic analysis to retail drug transactions to have been Mark Moore (1973). He pointed out that a would-be buyer of illicit drugs faces two "prices," one in money and the other in time, inconvenience, and risk in finding a seller. This second price, which Moore referred to as "search time" shares one of the primary properties of a money price (the higher it is, the smaller the volume consumed) but not the other (it is not a benefit to the seller and thus does not induce additional supply).

Higher search times have advantages over higher money prices as ways of reducing drug consumption: they neither enrich dealers nor impoverish users, and they reduce rather than (possibly) increasing the incentives for drug users to commit income- producing crimes such as theft, prostitution, and drug-selling. Search time is a probabilistic phenomenon, with a mean and a variance. The distribution of search times, which vary not only from drug to drug and from market to market but from buyer to buyer, depends on the number, geographic and social location, and aggressiveness of retail sellers. The number, location, and behavior of retailers are not directly influenced by changes in wholesale prices brought about by interdiction or high-level drug law enforcement. It is therefore primarily retail-level enforcement which influences search time, and restores or fails to restore order to neighborhoods disordered by retail drug markets. (Moore 1973, 1977; Kleiman 1988; Kleiman and Smith 1990)

In local markets as well as national ones, transactions (unwillingly) compete for scarce enforcement attention, thus creating external economies of scale. Further external economies are created by the fact that concentration of participants reduces search times. Thus retail markets tend to concentrate and to remain where they in the absence of large external forces. There may be a critical level of enforcement effort, continued over time, required to force a market below is minimum viable size. This strongly implies that retail enforcement efforts should be concentrated rather than being dispersed. (Kleiman 1988; Kleiman and Smith 1990; Caulkins 1990; Kleiman 1991)

Moore's original work on the subject of search times paid particular attention to undercover "buy-and-bust" operations as a means of increasing search time, particularly for new users, by making retail dealers fearful of dealing openly or with new customers. (Moore 1973, 1977) But the vocabulary of tactics directed at retail drug dealing is substantially larger than buy and bust alone.

Retail transactions can be divided into "flagrant" transactions, which take place in the open or in dedicated drug- dealing locations, and (relatively) discreet transactions which take place in multi-purpose indoor locations. Flagrant transactions are easier to interfere with, and create larger external burdens, than discreet transactions. (Kleiman and Young 1991) The last several years have seen the development of a large array of approaches for both officials and citizens to use to disrupt the conduct of flagrant retail drug markets (Burns and Conner 1991).

One approach to designing such tactics starts with an analysis of the conditions which make a retail drug market viable. It requires a place -- a venue -- in which buyers and sellers can meet to do business. The buyers need convenient access to that venue, the desire to buy the drugs for sale there, income to turn that desire into market demand, and a sense that they can buy with (at least relative) impunity. The sellers need operating scope within the venue, a supply of labor (their own or someone's else), attractive ways to spend or save the proceeds to maintain their incentive, a supply of drugs, and again a sense of impunity. Note that on this analysis all interdiction and high- level enforcement is encompassed within one factor (drug supply) of the ten.

Anything which interferes with any of these "factors of production" will reduce the volume of transactions. Tactics which do not involve arrests and trials (e.g., changing traffic patterns, noting license numbers of cars cruising to buy drugs and sending postcards to their owners, boarding up drug-dealing locations as public nuisances or for code violations) may not run into the system capacity constraints which have frustrated more conventional enforcement efforts. (Kleiman and Young, 1991; Press 1987)

All politics, it has been said, is local. Ultimately, all drug dealing is retail. The entire superstructure of international production, transport, distribution, and money laundering is supported by a base of retail sellers doing business with retail buyers. Learning how to use public authority and resources, and citizens' willingness to engage themselves, to reduce the volume of those transactions, particularly flagrant transactions, would seem to be the primary drug-policy challenge of the 1990s.

REFERENCES

Becker, Gary S., Michael Grossman, and Kevin M. Murphy. "Rational Addiction and the Effect of Price on Consumption." AEA Papers and Proceedings. *American Economic Review*, Vol. 81, No. 2, May 1991, p. 237-241.

Becker, Gary S., and Kevin M. Murphy. "A Theory of Rational Addition." *Journal of Political Economy*, 1988, Vol. 96, No. 4.

BOTEC Analysis Corporation. "Drug Abuse in Jackson County, Missouri: Problem Assessment and Recommendations." Cambridge: BOTEC, 1990.

Brown, George F., and Lester Silverman. "Retail Price of Heroin: Estimations and Applications," *Journal of the American Statistical Association*, Vol. 69, No. 347, September 1974.

Burns, Patrick, and Roger Conner. *Winnable War: A Community Guide to Eradicating Street Drug Markets*. Washington D.C.: American Alliance for Rights and Responsibilities, 1991.

Caulkins, J. P. "Distribution and Consumption of Illicit Drugs: Some Mathematical Models and Their Policy Implications." Ph.D. dissertation in Operations Research, Massachusetts Institute of Technology, 1990.

Cave, Jonathan, and Peter Reuter. *The Interdictor's Lot: A Dynamic Model of the Market for Drug Smuggling Services*. Santa Monica, CA: RAND Corporation, February 1988.

Howard, Ronald A. "On Making Life and Death Decision." In Schwing, R.C. and W.A. Albers, Jr., *Societal Risk Assessment*. General Motors Research Laboratories, 1980, pp. 89-113.

Huizinga, D.H., and D.S. Elliott "A Longitudinal Study of Delinquency and Drug Use in a National Sample of Youth: An Assessment of Casual Order." The National Youth Survey Project, Report No. 16, Boulder, CO: Behavioral Research Institute, 1981.

Kandel, Denise, Orta Simcha-Fagan, and Mark Davies, "Risk Factors for Delinquency and Illicit Drug Use From Adolescence to Young Adulthood." Journal of Drug Issues, Vol.16, No.1, (Winter 1986).

Kaplan, John. *The Hardest Drug: Heroin and Public Policy*. Chicago: University of Chicago Press, 1983.

Kelling, George, and James Q. Wilson. "Broken Windows: The Police and Neighborhood Safety." In *The Atlantic Monthly*, March 1982.

Kleiman, Mark A.R. "The Changing Face of Cocaine," Report to the Ford Foundation. *Cambridge MA.: Harvard University, John F. Kennedy School of Government, Program in Criminal Justice Policy and Management*, Working Paper #87-01-07, January 1987.

Kleiman, Mark A.R. "Crackdowns: The Effects of Intensive Enforcement on Retail Heroin Dealing." In *Street Level Enforcement: Examining the Issues*, edited by Marcia Chaiken. Issues and Practices in Criminal Justice. Washington D.C.: National Institute of Justice, 1988.

Kleiman, Mark A.R. *Marijuana: Costs of Abuse, Costs of Control.* New York: Greenwood Press, 1989.

Kleiman, Mark A.R. "Compliance and Enforcement in a Binary-Choice Framework." In *Modeling Drug Markets*. Cambridge MA.: Harvard University, John F. Kennedy School of Government, Program in Criminal Justice Policy and Management, 1991.

Kleiman, Mark A.R. *Against Excess: Drug Policy for Results*. New York: Basic Books, Forthcoming.

Kleiman, Mark A.R., Christopher E. Putala, Rebecca M. Young, and David P. Cavanagh. "Heroin Crackdowns in Two Massachusetts Cities." Office of the District Attorney for the Eastern District, Commonwealth of Massachusetts. Final report to the National Institute of Justice #85-JJ-CX-0027, 1988.

Kleiman, Mark A.R., and Kerry D. Smith. *State and Local Drug Enforcement: In Search of a Strategy*. Chicago: University of Chicago, 1990.

Kleiman, Mark A.R., Denise Kulawik and Sarah Chayes. "Program Evaluation: Santa Cruz Regional Street Drug Reduction Program." Cambridge, MA.: BOTEC Analysis, 1990.

Kleiman, Mark A.R., and Rebecca M. Young. "The Factors of Production in Retail Drug Dealing." Cambridge MA.: Harvard University, John F. Kennedy School of Government, Program in Criminal Justice Policy and Management, 1991.

Moore, Mark H. "Policies to Achieve Discrimination in the Effective Price of Heroin," *American Economic Review* 63, May 1973: 270-279.

Moore, Mark H. *Buy and Bust: The Effective Regulation of an Illicit Market of Heroin*. Lexington MA: Lexington Books, 1977.

Moore, Mark H. "Limiting Supplies of Drugs in Illicit Markets." *Journal of Drugs Issues*, Spring 1979, pp. 291-308.

Moore, Mark H. "An Analytic View of Drug Control Policies." Cambridge, MA.: Harvard University, John F. Kennedy School of Government, Program in Criminal Justice Policy and Management, 1990, Working Paper #90-01-19.

Moore, Mark H., and Dean Gerstein. *Alcohol and Public Policy: Beyond the Shadow of the Prohibition*. Washington D.C.: National Academy Press, 1981.

Moore, Mark H. and Mark A.R. Kleiman. "The Police and Drugs," *Perspectives on Policing*, No. 11, September 1989.

Morgan, John P. "Was Alcohol Prohibition Good for the Nation's Health?" New York: City University of New York Medical School, 1991, unpublished.

Musto, David F. *The American Disease: Origins of Narcotics Control.* London: Oxford University Press, 1987.

Musto, David F. "Opium, Cocaine and Marijuana in American History." *Scientific American*, July 1991, pp. 40-47.

National Household Survey: 1990. Washington D.C.: US Department of Health and Human Services.

The NNICC Report 1989: "The Supply of Illicit Drugs to the United States." Washington D.C.: National Narcotics Intelligence Consumers Committee, June 1990.

Press, Aric. "Piecing Together New York's Criminal Justice System: The Response to Crack." New York: New York Bar Association, 1987.

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

Reuter, Peter. "Quantity Illusions and Paradoxes of Drug Interdiction: Federal Intervention into Vice Police." In *Law and Contemporary Problems,* Winter, 1988.

Reuter, Peter, Gordon Crawford, and Jonathan Cave. *Sealing the Boarders: The Effects of Increased Military Participation in Drug Interdiction.* (Santa Monica, CA: RAND Corporation, 1988.

Reuter, Peter, and Mark Kleiman. "Risks and Prices: An Economic Analysis of Drug Enforcement." In *Crime and Justice: An Annual Review of Research,* Norval Morris and Michael Tonry eds. Chicago: University of Chicago Press, 1986.

Reuter, Peter, Robert MacCoun, and Patrick Murphy. "Money from Crime: A Study of the Economics of Drug Dealing in Washington, D.C." Washington D.C.: RAND, 1990, R-3894-RF.

Reuter, Peter and John Haaga, eds. "Improving Data for Federal Drug Decisions." RAND Note Washington D.C.: Bureau of Justice Statistics, 1991.

Rhodes, William, and Douglas McDonald. "What America's Users Spend on Illegal Drugs." Washington D.C.: Office of National Drug Control Policy, June 1991.

Schelling, Thomas C. *Micromotives and Macrobehavior.* New York: Norton 1978.

Schelling, Thomas C. *Choice and Consequence*. Cambridge, MA.: Harvard University Press, 1984.

Viscusi, Kip W. *Risk by Choice: Regulating Health and Safety in the Workplace.* Cambridge, MA: Harvard University Press, 1983.

Wharton Econometrics. "Anti-Drug Law Enforcement Efforts and Their Impact," Prepared for U.S. Customs Service. Bala Cynwyd, Pennsylvania: U.S. Department of the Treasury, 1987.

Zeckhauser, Richard, and John Pratt, eds., *Principal and Agent: Structures of Business*. Boston: Harvard Business School Press, 1991.

Zimmer, L. "Operation Pressure Point: The Disruption of Street- Level Drug Trade in New York's Lower East Side." Occasional paper. New York: New York University School of Law, Center for Research in Crime and Justice, 1987.

COMPLIANCE AND ENFORCEMENT IN A BINARY-CHOICE FRAMEWORK

Mark A.R. Kleiman

COMPLIANCE AND ENFORCEMENT IN A BINARY-CHOICE FRAMEWORK

Mark A.R. Kleiman

ABSTRACT

When many potential violators of a rule weigh the costs of compliance against the risks of punishment, a constraint on enforcement capacity will create an interdependence among their choices; the expected-value punishment for violation falls as the number of violators rises. Under simple assumptions, this will produce a situation with two stable equilibria -- near-perfect compliance and near-universal violation -- and an unstable crossover ("tipping") point between them. It is possible to design enforcement strategies to prevent crossover to widespread violation and to induce crossover to near-perfect compliance. In such circumstances, temporary enforcement surges may have lasting benefits. Singling out groups of violators or violation types for temporary intensive enforcement may be more effective than spreading enforcement effort equally across violations. This class of phenomena may help explain the time-track of drug prices and volumes and the geographic concentration of drug markets.

COMPLIANCE AND ENFORCEMENT IN A BINARY-CHOICE FRAMEWORK

Consider a group of risk-neutral, rationally self-interested decisionmakers, homogeneous with respect to underlying preferences, each of whom makes a series of binary choices between a permitted and a forbidden course of action. For example, imagine 1000 commuters to the same office building, each of whom chooses each day between parking in a private off-street lot (paying the parking fee) and parking in a free public lot intended for shoppers, with a posted two-hour limit.

Assume that the choices are not directly interdependent (each lot has more than 1000 spaces). Whether any one decisionmaker "complies" (parks in the private lot) or "violates" (stays in a free space) depends on his perception of the cost of compliance and the risk of punishment. Assume to start with that the decisionmakers absorb no psychic cost from violating and gain no psychic benefit from complying; those commuters are not conscientious about obeying parking regulations.

Let C be the additional cost of complying rather than violating (the parking fee, say $6). Then as long as C is positive, self-interested decisionmakers will choose to violate unless given an incentive to comply. Let P be the penalty for violation (the fine for a parking ticket, which we will assume is self-collecting once written). As long as P is greater than C, say $10, it pays to comply, and our rational decisionmakers will do so. If N is the number of decisionmakers, N_C the number complying, N_V the number violating, and $N_V/N = R_V$ the violation rate, then R_V will be close to zero if $P>C$. An occasional commuter may try parking for free to see if he actually gets a ticket; new commuters are likely to do so once or twice. But if each misparked car gets a ticket, personal experience will drive the commuters to the pay lot, and word-of-mouth information exchange will discourage experimentation by others. By the same token, if $C>P$ (the lot charges $12), experience and communication will eventually lead all decisionmakers to violate rather than comply, and R_V will tend toward unity.

Now let T be the quantity of enforcement action taken on a given day (the number of tickets written). T is the lesser of N_V and E, where E represents the capacity of the enforcement system (there is only one ticket-writer, who can write only 100 tickets per day).

The enforcement capacity constraint E introduces an interdependency among decisionmakers. If all the commuters could agree to mispark on the same day, then only 100 of them would get tickets. Assuming that the ticket-writer chooses at random which cars to tag, in the long run each commuter will pay an expected value penalty of:

$$E\,[P] = P * (E/N_V)$$

One hundred tickets at $10 divided by 1000 cars equals $1 per day to park. So it would pay each commuter to park illegally, even at a $6 lot fee, if all the rest were doing so. But if R_V is now close to zero, any one commuter would be ill-advised to test the system. If $P>C$ and $E * P/N < C$, we have the familiar tipping-point phenomenon (Schelling, 1972): two stable equilibria at $R_V = 0$ and $R_V = 1$, and an unstable crossover point, $\bar{R}_V$ at:

$$P * (E/N_V)\,C$$

$$\text{or, substituting, } (P * E)/(R_V * N) = C$$

$$\text{and thus, } \bar{R}_V = (P/C) * (E/N)$$

Thus, at the critical value $\bar{R}_V$, the violation rate is the product of the penalty-to-compliance-cost ratio and the enforcement-capacity-to-population ratio. If the lot costs \$6, a parking ticket is \$10, the commuter population 1000, and the ticket-writing capacity 100, then the tipping point is:

$$\bar{R}_V = (10/6) * (100/1000) = 1/6$$

If 167 commuters park illegally, the expected cost to each one is

$$(100/167) * \$10 \approx \$6$$

and they are roughly indifferent between the public lot and the pay lot.

Where multiple equilibria exist, history as well as underlying conditions help determine the outcome. Observing that one group of decisionmakers uniformly complies, while another group, facing identical values of P, C, and E/N, uniformly violates, one cannot conclude that the first group is socially responsible and the second incorrigible. There may be much less difference between them may be much less than meets the eye: an accident of history.

Arranging for the right set of historical accidents is a task for policymakers. In tipping-point situations, temporary interventions can have lasting effects. If the authorities wish to minimize violations (rather than maximizing ticket revenues), hiring a temporary additional ticket-writer when R_V is hovering around $\bar{R}_V$ will help tip things in the right direction. With a half-time ticket-writer added to the full-time one,

$$E[P] = (150/167) * 10 = \$9$$

at which point the pay lot is cheaper than misparking. Once R_V drops below .1, the part-time ticket-writer can be dispensed with, since the full-time ticket-writer can ticket all of the remaining violators. This approach will work even if the intervention is known in advance to be temporary, unless a large group of commuters cooperates in noncompliance and does so simultaneously. How long a "temporary" intervention might need to last depends on the information-gathering strategies of decisionmakers. (It may take some time for commuters to notice that the probability of getting a ticket has risen from 10% to 20%.)

Information can be deliberately manipulated. Fliers on the windshields of vehicles can tell violators that today's ticket is no accident. To the extent the fliers are believed, they will be self-confirming, since the probability of a ticket rises as the violation rate falls.

In the extreme, information efforts could even be substituted for action. If, near the tipping point, 40% of those who otherwise would have been violators believed a sign reading, "Beginning Monday, every illegally parked car will be ticketed," that statement would turn out to be true even if no part-time ticket-writer were hired. Only if the message is widely disbelieved will it turn out to be false.

When such information-only strategies succeed, they do so at low cost. But when they fail, and the absence of real effort behind the advertising becomes apparent, the damage may be enduring, in the form of lost credibility for future informational efforts and thus the need for longer temporary interventions.

Since systems do not hover long around unstable equilibria, instances where relatively small temporary interventions will be efficacious are atypical. How could the authorities turn around the situation starting from massive non-compliance, $R_V \approx 1$?

One answer is massive temporary help. To make violation unprofitable, we need:

$$P * (E/N_V) > C$$

Call this required level $\bar{E}$.

At $R_V = 1$, this condition is equivalent to:

$$P * (E/N) > C$$

$$\text{or } (E/N) > (C/P)$$

that is, the enforcement-capacity-to-population-ratio must exceed the cost-of-compliance-to-penalty-ratio. In our example, we need E such that:

$$(E/1000) > 6/10$$

$$\text{or } E > 600$$

which requires 6+ ticket-writers. Presumably, if E is just above $\bar{E}$, the process of readjustment will be slower than if we start with E closer to N (10 ticket-writers in this instance). With some data about behavior, it would be possible to find the time-path of E that brings about the transition from $R_V \approx 1$ to $R_V \approx 0$ using minimum resources (ticket-writer-days), and to design information strategies to reduce this number further.

But not all hope is lost even if the enforcement capacity constraint is inflexible, if only the homogeneity of violators (or violations) is less than perfect. If, for example, the public lot is a rectangular array with 100 spaces per line, we can assign the sole ticket-writer to line #1 (with or without publicity, though "with" will work faster). Now parking in line #1 is worse than paying to park. Once decision makers understand this, no one will park in line #1. Next announce 100% effort for line #2 as well as line #1, and repeat. As long as the lot is finite in size, eventually the area not under perfect enforcement will be less than 1000 cars, and some would-be violators will be squeezed back to the pay lot. (Anyone who tries to "squeeze" back toward the "clean" areas gets ticketed.) Eventually, the number of non-"clean" spaces drops below R_V, and the system tips toward $R_V = 0$. (See Sloan-Howitt and Kelling, on the process of discouraging graffiti-writing on the New York subways.)

Spatial discrimination in enforcement is easy to visualize, but any kind of discrimination that picks out no more than $\bar{R}_V$ violators at a time will serve. One could start with cars with plate numbers ending in zero and work through the digits. One could start with blue cars and work through the colors (breaking up white cars into sedans, hatchbacks, hardtops, and sports cars). One could start with the current model year and work backwards (dividing recent years by make and model). Even if the basis of discrimination is not deducible by the violators, ten consecutive tickets should persuade any reasonable violator that the ticket-writer is making him a special target and that, for him, the pay lot is effectively cheaper.

All this ingenuity is for naught, however, if $C>P$. If the price of paid parking goes to $12, no enforcement effort will keep R_V from heading toward unity. Once that has happened, simply raising P will not suffice to restore compliance; increased enforcement activity will be needed as well.

This analysis might be extended and enriched in a number of directions. Here it will suffice to indicate three of them.

EXTENSIONS

Heterogeneity among Violators

The assumption that potential violators are homogeneous with respect to their costs of compliance and their subjective evaluation of the penalties imposed is quite restrictive. Note that if a relatively small number of potential violators are particularly inclined to violate rather than complying -- because their cost of compliance is high, their sensitivity to the punishment is low, or their information-processing styles lead them to respond only slowly to evidence that the chance of being penalized has grown -- they may help "tip" the entire system over into a region in which high violation rates are self-sustaining.

From a policy-maker's perspective, this raises the question of programs to single out frequent violators for special attention: help in reducing their compliance costs; special penalties; directed enforcement or information campaigns to increase their perceived non-compliance costs; or measures to "incapacitate" them (i.e., to make it physically impossible for them to violate; in the example, by impounding their vehicles). Any such program may run into complaints about equal protection, particularly if the process by which high-rate offenders are identified is imperfect. This is one analytic approach to the problem of "dangerous offenders." (See Moore, Estrich, McGillis, and Spelman, 1984)

Varying Severity Rather than Probability

If all potential violators are risk-neutral, their response to a given combination of punishment probability and punishment severity will depend only on the product of the two, the expected value of punishment per offense. In this case, increased severity is a perfect substitute for increased probability. Increasing severity rather than probability will be a particularly attractive solution where the deadweight costs of punishment are small and the capacity to punish is not scarce.

But if potential violators are risk-averse in welfare terms (if, that is, making an actuarially fair bet -- increasing the variance of their wealth without changing its expected value -- reduces their expected utility) then a low-probability, high- severity punishment regime will produce a deadweight loss in utility terms even if the penalty itself is a costless transfer. Imposing a million dollars in parking fines could be more easily accomplished by issuing a thousand tickets for a thousand dollars each than by issuing one hundred thousand tickets for ten dollars each, but the utility loss to the recipient of a thousand-dollar ticket is almost certainly more than one hundred times the utility loss to the recipient of a ten-dollar ticket. This consideration gives analytical backing to the intuition that such fines would be unfair.

The position is even worse if potential violators are behaviorally risk-seeking (for the kinds of reasons explored by Kahnemann and Tversky 1984, because large infrequent fines would be less effective deterrents as well as producing greater utility losses to those punished, when compared with small infrequent fines of the same aggregate value.

The Deadweight Costs of Punishment

A fine is close to a pure transfer. Aside from the ticket-writer's wage, the resource cost of moving money from the offender's wallet to public fisc is perhaps no greater than the collection costs for other forms of municipal revenue-raising. We can thus be largely indifferent about the number of enforcement actions taken.

But this is a very special case. Resource-guzzling adjudication systems and non-pecuniary penalties create large deadweight losses (in the U.S., about $15 billion per year to operate prisons and jails, plus whatever 1 million prisoners and their intimates could pay for their freedom). Under these circumstances, reducing the total volume of punishment ($T * R_V$) is a policy objective. This effect increases the social cost of making the transitions from high-violation to low-violation equilibria.

IMPLICATIONS

Two general observations, one theoretical and one practical, may be drawn from the analysis above. On a theoretical level, it suggests that the attempt to explain differences in offense rates across social settings by examining current conditions and the current dispositions of potential offenders may by misdirected. On a practical level, it suggests that a mechanical application of the principle of horizontal equity to the problem of crime control may be extremely costly.

Social phenomena, like physical systems, can usefully be divided into two classes according to the dependence (*vel non*) of their equilibrium or resting states on their initial conditions. Systems with a single stable equilibrium will tend toward that condition from any set of initial values; these systems are sometimes called "thermodynamic." If two similar systems of that type reach different resting states, it can only be concluded that they differ in the parameters which determine the (unique) equilibrium, rather than in initial conditions. If, for example, two bodies of water saturate at different salt concentrations, they must have different temperatures or concentrations of other solutes; previous conditions (e.g., the fact that one started out cold and the other hot, or that in one case the salt started out in solid form, while the other started out as brine) cannot explain differences in equilibrium. If two solutions have saturated at different concentrations because of a difference in temperature, equalizing the temperatures will (after some delay) equalize the concentrations. Thermodynamic systems have no memories.

By contrast, a soufflé is a non-thermodynamic system. Its current condition is in part determined by its history; once it has fallen, no manipulation of external conditions will cause it to rise again. The investigator will search the present in vain for reasons why one soufflé and not another has fallen; the answer is in the past.

The analysis above suggests that some features of the task of enforcement make it resemble cooking rather than high-school chemistry. History plays a part, along with current conditions, in determining offense rates. This suggests that much of the criminological literature about the (present) causes of differing offense rates across social settings may be directed at answering the wrong question. High and low offense rates may be self- sustaining, and the assumption that there must be some difference in current conditions or dispositions between a relatively law-abiding population and a relatively offense-prone population may not be justified.

(In one sense, of course, the current ratio of punishments to offenses constitutes data about the present rather than about the past. But any explicit attempt to include that ratio in an explanatory schema runs into almost insuperable problems of specification, because it puts the offense rate on both sides of the explanatory equation. This gives rise to the difficulty of empirical deterrence research, as outlined in Blumstein, 1978.

The observation that crime may not be thermodynamically determined gains interest in light of the difficulty criminologists have had in explaining the differences in offense and victimization rates across ethnic groups by differences in the other (present) characteristics of such groups. (See, e.g., Wilson and Herrnstein, 1985. The same observation could be applied to cross-national analysis. In short, the assumption that those who commit many crimes are either constitutionally or situationally crime-prone may not be justified.

On a practical level, the notion of recovering from a high-offense situation by concentrating enforcement resources on a single class of potential violators until they are pushed over the "tipping point" seems to be in direct conflict with the idea that criminal justice agencies should respect horizontal equity -- the principle which requires that like cases be treated alike -- among offenders. The counter-principle of "divide and conquer," while it has (on the assumptions of this analysis) considerable practical utility, has no equivalent moral standing.

Unless the discrimination among offenders is actuated by animus against one or another social subgroup, the violation of horizontal equity is more apparent than real. The sort of enforcement considered here is always probabilistic, and we can think of a (random) decision to single out the drivers with license plates ending in zero as simply substituting a two-stage lottery for a one-stage lottery in determining who gets the tickets. But where the stakes are any larger than a parking ticket, the apparent unfairness will be disturbing to many and intolerable to some.

This would not be the only instance in which the operational requirements of crime control conflicted with the demands of justice. Rather than pretending that no such conflict exists, or vainly attempting to prove that one or the other deserves an absolute preference, it might be wiser simply to admit the possibility of a tension and to look for enforcement strategies, and justifications for enforcement strategies, which sacrifice as little of each objective as possible to the demands of the other.

APPLICATIONS

The potential applications of this analysis are as varied as the situations in which a rule is to be enforced. In addition to a multitude of private situations (securing compliance with administrative deadlines, encouraging one's children to do household chores, noise control within an apartment complex), one could apply it to compliance with taxation, water pollution laws, or workplace safety regulations.

This model also seems to help elucidate the logical structure of rioting, whether the "topic" of the riot is race, religion, famine, politics, or football. Many persons who would not find the material-plus-psychic rewards of breaking a shop window and stealing the contents an adequate inducement to endure the level of enforcement risk ordinarily associated with commercial burglary can be induced to participate in such activities once many others are doing so and thus competing for a limited supply of enforcement attention. These might be value in an analysis of riot control tactics from this point of view.

The applications to illicit markets, and particularly the markets in prohibited drugs, are of particular interest and complexity. This model helps explain both the time-patterns of drug industry volumes over time and the geographic distribution of local drug market activity.

Positive Feedback in Drug Volume and Prices

In a simple comparative-statics model, the aggregate revenues in the market for any drug will equal the aggregate costs, including the costs of enduring or avoiding enforcement. (Reuter and Kleiman 1986) In such a model, aggregate volume is determined by the (cost-driven) price and the demand curve; the market will equilibrate where the consumer of the marginal dose is exactly willing to pay the marginal cost of delivering that dose through the barriers erected by enforcement. Except in the implausible case where demand is completely inelastic to price, volume will fall as the price rises.

If the aggregate level of enforcement-imposed costs depended only on the level of enforcement activity (i.e., if drug enforcement were strictly analogous to the stylized model of parking enforcement given above), and if the level of enforcement activity were fixed in the short run, then the drug markets would be characterized by industry-wide economies of scale: the larger the volume, the lower the price, as transactions protected one another by competing for enforcement attention. This effect, added to any price-elasticity of demand, would create a positive- feedback loop in the physical volume traded in any drug market. An increase in market activity in one period would tend to decrease prices in that period, as a larger number of transactions shared among themselves a fixed burden of enforcement-generated costs. Increased consumption due to lower prices in that period, would, to the extent that drug- taking is habitual behavior (and thus consumption in period n is a complement to consumption in period n + 1) lead to increased demand in the next period. Insofar as dealers correctly anticipated that growing market and attempted to deliver more drugs in the subsequent period, prices in the subsequent period would again be lower due to increased volume, and so on. The same process would work in reverse once a market began to shrink. (Kleiman 1987)

This model is a limiting case because it makes two extreme assumptions: that the costs imposed by enforcement depend only on the level of enforcement activity (rather than rising with the level of market activity as well) and that the allocation of enforcement resources is independent of the volume of drug distribution. Neither assumption is likely to be strictly true. A given number of agents are likely to make a larger aggregate volume of seizures of drugs and assets, and more (quality- adjusted) arrests in an active, particularly a growing, market than in a quiet and stagnant one, both because much of the activity of agents is search behavior -- search time in general falls as the density of the searched-for entities rises -- and because a larger volume of genuine illicit activity provides camouflage for undercover activity. Consequently, agent's incentives will lead them toward active drugs and active regions, thus somewhat proportioning the volume of enforcement activity to the volume of illicit transactions. Moreover, policy-makers, responding to public demand, are likely to back this natural tendency with explicit orders. (On the other hand, when the limiting capacities are court time and prison space rather than agent-hours, the real case may approach the limiting case rather closely; see, for example, the description of cocaine enforcement in New York City by Aric Press, (1987).

However, the positive-feedback tendency will remain as long as the growth in enforcement-driven costs due to greater enforcement productivity and increased enforcement resources is slower than the growth in physical drug volume. This will create an apparent paradox: the drug market will continue to grow despite growth in measured enforcement outputs (arrests, seizures, convictions). Thus, during the growth phase of a drug market, enforcement will appear entirely ineffectual in controlling the illicit industry.

It is not far-fetched to regard the collapse of cocaine prices and explosion in cocaine consumption during the late 1970s and 1980s as an illustration of this positive-feedback phenomenon. That is a cheerful reflection as the evidence mounts that cocaine consumption has peaked, because it suggests the possibility of a comparably dramatic consumption collapse and price rise during the 1990s.

Geographic Concentration

One reason for the geographic concentration of drug markets is that it minimizes search costs for buyers and sellers. This analysis provides an additional reason in the form of enforcement costs.

Fish swim in schools and ruminants in flocks because numbers provide relative safety from predators; an isolated prey individual is far more likely to be eaten than that same individual would be surrounded by hundreds of its species-mates. The sole drug dealer in a neighborhood is far more likely to be the personal focus of police attention than one dealer in a crowd. All that is required for this to be the case is that enforcement activity be less concentrated geographically than illicit activity.

Policy Implications in Drug Enforcement

This analysis points toward the virtues of rapid response and concentration in drug law enforcement.

Once it appears that the market for a given drug has begun to grow, problems can be averted and resources saved in the long run by rapidly mobilizing additional enforcement response so as to make aggregate enforcement-imposed costs rise more quickly than physical volume. Cocaine enforcement signally failed to accomplish this objective through the first dozen years of the cocaine epidemic. Growing evidence of an incipient take-off in the heroin market may therefore warrant the prompt attention of policy-makers.

At the local level, the implication of this model is that concentrating attention on one or a few market areas is likely to have larger payoffs than dispersing attention across all the markets in a city or region. While good fisheries management consists in restricting the fish catch below its sustainable maximum, good enforcement management -- where the "catch" of arrests counts as a cost rather than a benefit -- may consist in deliberately "overfishing" in a few places. In a city with ten market areas, each of which can easily absorb one-tenth of the available enforcement capacity, no one area may be able to sustain the concentration of all or almost all of the available resources for a period of months, after which a lower residual enforcement effort may suffice to keep a closed market closed.

This analysis also suggests that displacement may be a smaller problem for local drug law enforcement than would otherwise appear to be the case. If local drug dealing is like the parking example, local markets have a minimum viable size determined by the background level of enforcement activity. (Caulkins 1990) Unless buyers and sellers have a coordination mechanism which allows the bulk of them all to move to the same new location, displaced transactions will also be dispersed, and will therefore face higher enforcement costs than they did when they were concentrated.

Here again, the key to preventing the growth of replacement markets would seem to be rapid response to early signs of activity in new areas. Such rapid response would require some form of information-gathering: attention to reports from patrol officers, analysis of spontaneous citizen complaints, encouragement of such complaints through publicized "tip lines," interrogation of drug users subsequent to arrest (particularly on non-drug charges), or collation of information from street ethnography or the treatment system.

REFERENCES

Blumstein, A., J. Cohen, and D. Nagin, eds. *Deferrence and Incapacitation: Estimating the Effects of Criminal Sanctions on Crime Rates.* Washington, DC: National Academy of Sciences, 1978.

Caulkins, J.P. "Distribution and Consumption of Illicit Drugs: Some Mathematical Models and Their Policy Implications." Ph.D. dissertation in Operations Research, Massachusetts Institute of Technology, 1990.

Kahnemann, Daniel, Paul Slovik, and Amos Tversky, ed. *Judgement Under Uncertainty: Heuristics and Biases* Cambridge, MA: Harvard University Press, 1984.

Moore, Mark H., Susan R. Estrich, Daniel McGillis, and William Spelman. *Dangerous Offenders: The Elusive Target of Justice* Cambridge, MA: Harvard University Press, 1984.

Press, Aric. "Piecing Together New York's Criminal Justice System: The Response to Crack." New York, New York Bar Association, 1987.

Reuter, Peter, and Mark Kleiman. "Risks and Prices: An Economic Analysis of Drug Enforcement." In *In Crime and Justice: An Annual Review of Research*, Vol. 7, edited by Michael Tonry and Norval Morris, 289-340. Chicago: University of Chicago Press, 1986.

Schelling, Thomas C. "On the Ecology of Micromotives." *The Public Interest.* Reprinted in Schilling, *Micromotives and Macrobehavior.* New York: Norton. 1978.

Sloan-Howitt, Maryalice, and George L. Kelling. "Subway Graffiti in New York City: 'Gettin Up' vs. 'Meanin It and Cleanin It'." In *Security Journal*, 1990, Vol. 1, No. 3, 1990.

Wilson, James Q., and Richard Herrnstein. *Crime and Human Nature.* New York. Simon & Schuster, 1985.

THE FACTORS OF PRODUCTION IN RETAIL DRUG DEALING

Mark A.R. Kleiman and Rebecca M. Young

THE FACTORS OF PRODUCTION IN RETAIL DRUG DEALING

Mark A. R. Kleiman and Rebecca M. Young

Abstract

Drug markets and drug control have traditionally been analyzed single-dimensional continua: e.g., supply vs. demand, high-level vs. low-level enforcement. This paper uses the economic analysis of factors of production to sketch a fuller model of the conditions which support drug consumption and distribution. The 10 factors of production necessary for drug markets are a common venue; buyers' access to the venue, their desire for drugs, income to spend on drugs and their sense that they can buy with impunity; sellers' operating scope within the venue, supply of drugs, ways to spend or save money earned, supply of labor, and sense that they can sell with impunity. Drug control policies can be designed to attack each of these contributors to the operation of retail drug markets, particularly "flagrant" markets which operate in the open or in locations dedicated to drug dealing (such as "crack houses").

THE FACTORS OF PRODUCTION IN RETAIL DRUG DEALING

The past decade has seen an explosion both of certain kinds of drug dealing and of public concern about drug dealing and drug consumption. But neither the nature of the problem nor the range of alternative approaches to controlling it is well understood.

Public policies to deal with the drug problem are often dichotomized into supply side approaches (law enforcement) and demand side approaches (education and treatment). That distinction has been persuasively criticized as overly simplistic (Moore 1979; Office of National Drug Control Policy 1989).

Both Moore and the authors of the National Drug Control Strategy argue that a sharp distinction between supply and demand side cannot be maintained in light of the economics of illicit drugs. Moore illustrates this point by reasoning that drug dealers who successfully undergo drug treatment (a demand side strategy) reduce the supply of drug dealing labor, thus reducing the supply of drugs. Nonetheless, the supply-demand explanation still plays a powerful role in public discussions of drug policy, so much so that the ratio of "supply side" to "demand side" efforts in the federal budget has become a matter of ideological and even partisan debate (Majority Staffs of the Senate Judiciary Committee and the International Narcotics Control Caucus 1990). In this paper, we maintain the distinction for taxonomic purposes, and attempt to provide a model that accommodates for the ways in which the basic distinction oversimplifies.

Another, less publicly resonant, debate concerns the appropriate allocation of enforcement resources along the chain of production and distribution from raw crops in the field to the final retail sale. Here the argument is between "high-level" approaches (crop eradication and laboratory destruction, source country enforcement, interdiction, and investigation of large-scale domestic traffickers) and "low-level" approaches aimed at disrupting retail trafficking (Chaiken 1988; Kleiman and Smith 1990; Moore 1973; Reuter and Kleiman 1986).

Some argue that these drug-specific policies are largely irrelevant, and that the drug problem is caused by deeper economic and social problems and can only be cured by broader economic and social reforms (Currie 1985). The discussion on this point resembles, in its form, its ideological loading, and its ultimate futility, the debate on the "root causes" of crime (compare Wilson 1983). This paper will restrict itself to discussion of more proximate causes and cures.

Alongside the pragmatic debate about how best to apply resources to diminish the extent of the problem, there is a social-scientific debate about how best to describe and explain the phenomena. Some neighborhoods, some cities, some countries are home to far more, and far more flagrant, retail drug dealing than others. What explains the variations? Which of the important explanatory variables can be manipulated by deliberate public action?

It seems unlikely that only two dimensions (supply vs. demand, high-level vs. low-level) are adequate for either prescription or description. The conceptual model they represent is simply not rich enough. This paper attempts to sketch a fuller model of the conditions which support drug consumption and distribution and to illustrate policies which specifically target each of those conditions.

The supply vs. demand debate reflects the extent to which both academic and official thinking about the drug situation has been influenced by the language and explanatory schemata of economics. Over the past two decades, there has been an effort to develop descriptions of drug-related behavior and approaches to drug abuse control based on a market metaphor (Caulkins 1990; Moore 1973, 1976; Kleiman 1989; Kleiman, Lawrence and Saiger 1987; Kleiman and Smith 1990; Reuter and Kleiman 1986; Spence

1977). Drugs are considered as having buyers and sellers, supply and demand curves, production functions, and industry structures. The enforcement-oriented literature tends to take the demand function to be fixed (i.e., given externally) and deals with various policies to change the supply curve. In thinking about prevention and treatment, the supply situation (prices, non-price "costs" of purchase, and search times) is assumed to be fixed, and the focus is on changing demand. This perspective provides the basis for the conceptual division of programs into demand side and supply side, and the identification of supply with enforcement and demand with prevention and treatment. It was David Ricardo in the nineteenth century who first analyzed the economics of agriculture in terms of three factors of production: land, labor, and capital (Heilbroner 1953; Ricardo 1951). The concept of factors of production is straightforward: It is possible to list the items required to produce any commodity, to ask for any given production technology how much more of the commodity could be produced with one additional unit of input (say an extra hour of labor) holding the other inputs fixed, and to distinguish among different production technologies based on the mix of factors used (handicraft tends to be labor-intensive, automated production capital-intensive).

The study of production was first developed for agriculture and manufacturing. Several adaptations were required to make it applicable to retail sales and service industries, which make up a growing share of economic activity. From the production-management viewpoint, the distinguishing fact about such activities is that the customer is part of the production process. (See Czepiel 1985)

The current paper is an attempt to identify the factors of production in the process that produces retail drug transactions, as an aid both to understanding the way that drug markets work and to understanding current drug control efforts and developing new approaches. From this perspective, one can ask about any anti-drug action, "What factor of production does it attack? Does it do so effectively? How will reducing the availability of that factor affect the production process?" (Of course, effectiveness makes up only half the picture when it comes to choosing policies. The other half is costs: in money, in alternative uses of the same resources, and in intrusion and inconvenience.)

For any fixed production technology (where a technology is defined by its ratios of factor inputs) and any given budget of inputs, there will be one (or sometimes more than one) factor in short supply relative to the others; this is called the "limiting factor of production."

For example, consider baking pound cake, where the recipe calls for a pound of sugar, a pound of butter, a pound of flour, and six eggs. If we have two pounds each of butter, sugar, and flour, but only half a dozen eggs, eggs are in relative scarcity and constitute a limiting factor. Adding to the butter supply will not allow us to produce any more pound cake (and subtracting from it will not restrict us to producing any less). But with an additional half-dozen eggs, we could produce two cakes instead of one; with three fewer eggs, we could only produce half a pound cake.

In the more general case where production technology is flexible rather than fixed (i.e., where it is possible to produce slightly more of the output by slightly increasing any one of the inputs), the question "What is the scarce factor of production?" is generalized to "For each factor of production, what is the marginal product?", that is, the ratio of additional output to additional input of a given factor, holding the other factors constant (the partial derivative of the output to a given input).

There is a branch of empirical economics devoted to deriving from operating data--primarily historical records of the quantities of inputs used and the quantities of outputs produced--what are called production functions: equations which relate output to inputs.

Similarly, if drug control policymakers identify the scarce or high-marginal-product factors of production in producing retail drug transactions, they may be able to bring their resources to bear at points of maximum effectiveness (compare Moore 1979).

The power of this approach is limited by the difficulty of gathering data about drug markets, and the rapidity with which such markets grow, change, and fade. The more factors of production there are in a production function model, the richer the data set required to specify it empirically. Unfortunately, studies of local drug dealing are characterized by an extraordinary paucity of data. For that reason, any discussion of the factors of production in retail drug dealing must remain at the level of metaphor, rather than being elevated to the realm of exponents and residual terms. We can only hope than an unspecifiable model is more helpful than no model at all.

Consider an individual retail drug transaction: two persons meet and exchange contraband drugs for money. This requires a seller (labor), a buyer (customers), and a place for them to meet (venue). It also requires drugs and money: the seller needs a source of supply and opportunities to spend or save in order to prevent his incentive from flagging, and the buyer needs a source of income and a desire for the drug. Customers require access (physical and social) to a dealing venue. Buyers also need some chance of consummating that and similar transactions with impunity.

Thus, the most basic framework includes six factors of production--drugs, money, labor, customers, venue, and impunity. If we count sources and uses of money (i.e., income and incentive), and the access and desire of the customers, separately, the total number of factors grows to eight.

To make the same list differently and in greater detail: the market has buyers and sellers. They need a venue in common. The buyers need access to that venue, desire for drugs, income with which to buy them, and some chance of impunity. The sellers need operating scope within that venue, a supply of drugs, ways to spend or save the money they earn as dealers to maintain their incentive, labor (their own or that of others), and again impunity. Thus if we count buyers' and sellers' impunity separately, we have a total of 10 factors of production.

Table 1: The Ten Factors of Production in Retail Drug Dealing

Venue

Buyer's-Side Factors	Sellers-Side Factors
Access to Venue	Operating Scope
Desire for Drugs	Drug Supply
Income	Incentive
Buyers' Impunity	Labor
Sellers' Impunity	

Any program to shape the market, whether aimed at reducing consumption, or commerce, or the unwanted side effects of either, must target one or more of these six or eight or ten factors. Determining the right mix of such efforts is essential to defining a sensible strategy. But the major polarity within popular and political views of the subject is defined by varying degrees of emphasis on only three factors: drugs, sellers' impunity (both identified with law enforcement, with a "hard line" on drugs, and with political conservatism), and the desire of the consumers (identified with education and treatment, with "compassion," and with political liberalism). This focus leads to confusion, and in particular, to an inadequately differentiated view of the means and ends of drug law enforcement. (For a discussion of many of these tactics see Burns and Conner, 1991; for an application of this analysis to the drug markets in a metropolitan county see BOTEC Analysis Corporation 1990.)

The balance of this paper will list and briefly discuss some of the tactics which might be employed against each factor of production starting with venue and then addressing buyers'-side and sellers'-side factors.

ATTACKS ON VENUE

Venue goes far toward determining the other characteristics of a market. Venues are either flagrant or discreet, A streetcorner drug market and crack house are both indiscreet, one because it is in plain view, the other because of the lack of legitimate activity as "cover." A customer's living room is about as discreet as one can get, but candy-store owners, taxi drivers, elevator operators in hotels and office buildings, and bartenders can also deal discreetly because they have frequent legitimate occasion to be alone together with strangers.

Venues may also be categorized as outdoor or indoor. Outdoor venues cannot be eliminated--there will always be parks and streetcorners--but which of them are used can be affected by targeting other factors of production: sellers' operating scope, buyers' or sellers' impunity, or buyers' access.

When it comes to indoor venues, actions by public officials and private citizens can attempt to eliminate them entirely. Redesigning apartment buildings, particularly publicly owned ones, to eliminate interior corridors and boarding up or demolishing vacant buildings, destroy the spaces themselves. Locking external doors of apartment buildings and evicting, or refusing to rent to, known dealers, are ways of eliminating venues by keeping sellers away from them. Landlords can also be held responsible for ensuring that their buildings do not serve as venues and can be threatened with loss of the property (through the forfeiture process) or loss of its income stream (if the building is closed for fire, housing, zoning, or health code violations) for failure to take appropriate steps in that direction.

The more discreet the dealing venue, the less the access, particularly for novice customers. The less discreet the dealing, the greater the burden per transaction on nearby persons and institutions, both because flagrancy increases visible disorder and because obvious buyers and sellers are easy targets for violence and therefore likely to arm themselves and thus to be sources of violence as well.

Since both the law of search and seizure and the difficulty of direct observation make indoor dealing harder to enforce against than outdoor dealing, it is easier for sellers to find impunity indoors than outdoors. If enough potential indoor venues can be eliminated, the result may be fewer consummated transactions. On the other hand, to the extent that displaced indoor markets move outdoors, the neighborhood effects may be worse.

ATTACKS ON BUYERS'-SIDE FACTORS

Table 2: Tactics to Attack Buyers'-Side Factors

Access to Venue
- Parking and Traffic Enforcement
- Blocking off streets or making them one-way
- Checkpoints
- Doormen in housing projects
- Residents' use of picket signs and bullhorns

Desire for Drugs
- Anti-drug messages (school, media, neighborhood)
- Treatment (including maintenance)
- T.A.S.C.

Income
- Target hardening for property-crime targets
- Restitution orders for drug-involved offenders
- Anti-prostituion efforts
- Drug testing for employees and benefit recipients
- High-level enforcement

Buyers' Impunity
- "Sell and bust"
- Sales of "turkey dope"
- Observation arrests
- Questioning suspected market participants
- Residents' use of picket signs and bullhorns
- Seizure and forfeiture of vehicles
- Car checks and postcards
- Drug testing for drug-involved offerders
- Fines
- Publicity

Access

Buyers need convenient access to the drug-dealing venue. Anything which makes their access difficult, unpleasant, or risky will tend to reduce the volume of retail transactions.

Rigorous parking and traffic enforcement near open drug markets can serve to limit buyers' access to those locations. Blocking streets off or making them one-way can also reduce the amount of traffic in drug dealing areas. Some jurisdictions use traffic checkpoints, where motorists are asked to show license and registration, to deter drug buyers, particularly suburbanites who drive into urban neighborhoods to "score." Housing projects that are well-locked and that employ doorkeepers to verify the identity of all entrants can greatly reduce access to those locations by potential drug buyers. Residents of neighborhoods with active markets can make customers feel unwelcome by picketing, marching, shouting through bullhorns, taking photographs, and so on.

These tactics have a common advantage: they make sparing use, or no use at all, of the formal mechanisms of law enforcement and criminal justice tactics which discourage deals without making arrests are far less expensive than those which require arrests (and subsequent trials and punishments) to be effective. (See Press 1987.)

Desire

Reductions in buyers' desire for drugs directly affect the market by decreasing demand. Various educational efforts in schools, neighborhoods, and through the media can convince prospective drug users not to start (Pentz et al. 1989). There has been increasing use of uniformed police officers in schools nationwide. The most publicized such program, D.A.R.E. (Drug Abuse Resistance Education), originated in Los Angeles, has been replicated in many other locales, and appears to be mildly effective in reducing students' use of illicit substances (Evaluation and Training Institute 1988).

Drug treatment is another way to attack the buyers' desire for drugs. There is clear evidence that heroin users in methadone maintenance programs buy less heroin (and commit fewer crimes) than addicts not in such programs, and that residential treatment programs (whether participation is voluntary or coerced) can cause long-term reductions in drug-buying (Anglin and McGlothlin 1981; Anglin and Speckart 1986, 1988; Speckart and Anglin 1986). T.A.S.C. (Treatment Alternatives to Street Crime) programs try to reduce the desire for drugs on the part of drug-using offenders by requiring them to accept drug treatment as an alternative to prison.

Income

Many strategies can be employed to limit the income of drug buyers. Target hardening for property-crime targets (e.g., unstealable car radios) helps to eliminate income sources of drug-using property criminals who sell their loot and use the profits to buy controlled substances. Requiring drug-involved offenders to pay restitution to their victims limits their disposable income. Antiprostitution efforts, to the extent that they are successful, decrease prostitutes' earnings and thus their available cash for drug purchases. Drug testing employees and terminating those with dirty urines who refuse treatment clearly cuts down on their ability to buy drugs. Similar strategies can be employed with recipients of income support payments and public housing residents.

Anything that increases the price of drugs helps put a strain on buyers' incomes. This is one of the goals of high-level drug enforcement (the other being to limit drug supplies). If buyers respond to higher prices by reducing drug consumption, that is all to the good. If, on the other hand, they respond by cutting back on food, clothing, and shelter, or by increasing their income from illicit sources, successful high-level enforcement may have perverse effects (Brown and Silverman 1974).

Buyers' Impunity

The more reason buyers have to fear detection and arrest and the more harassment they experience, the less buying they are likely to do. Strategies aimed at buyers' impunity include a variety of law enforcement tactics. "Sell and bust" operations using undercover police officers posing as drug dealers can be effective in deterring novice customers.

Undercover officers selling inert substances packaged as drugs ("turkey dope") discourage users who discover they have wasted good money on bad drugs. Police officers who observe drug transactions can make arrests on the basis of their observations, thus reducing buyers' (and sellers') impunity in street markets. Even stopping suspected drug market participants for questioning can serve as a deterrent.

Since there are more buyers of drugs than sellers, deterring drug-buying needs to be a high-volume activity. The criminal law is so expensive and capacity-limited that aiming criminal sanctions at buyers is difficult. This puts a premium on tactics which do not involve arrest and trial, such as the seizure and (administrative) forfeiture of vehicles driven by drug buyers.

An even simpler tactic aimed at buyers with wheels is to notify them that their presence has been recorded. Civilians and/or police officers can note the license plate numbers of out-of-neighborhood cars driving through drug market areas. Police can then send a postcard to the residence where the car is registered, saying that the car was seen at X location, known to be a place where illicit drugs are sold, and warning the owner that the vehicle is subject to forfeiture if drugs are found in it. (If the buyers are, for example, the children of the registered vehicle owners, the repercussions of such postcards may be substantial.)

To the extent that the criminal justice system imposes urine monitoring for drug use on persons on bail, probation, or parole, and imposes sanctions for positive tests, it should be able to cut down on the purchase activity of an important subset of buyers (Kleiman and Smith 1990; Toborg and Kirby 1984).

ATTACKS ON SELLERS'-SIDE FACTORS

Table 3: Tactics to Attack Sellers'-Side Factors

Operating Scope

- Towing cars
- Cutting brush
- Lighting
- Doormen in housing projects
- Anti-gun measures (seizures and gun control)

Drug Suppply

- Crop eradication
- Source-country law enforcement
- Interdiction
- Long-term undercover operations
- Historical investigation of drug conspiracies
- Electronic surveillance
- Financial investigation
- "Mr. Big" enforcement
- "Buy and bust" operations
- "Working up the chain" investigations
- Control of intermediate chemicals and diluents

Incentive

Denying Saving Opportunity:

- Money-laundering investigations
- Seizures and forfeitures
- Civil suits
- Fines

Denying Spending Opportunity:

- Tax investigation
- Seizure of displayed wealth as evidence of dealing
- Checking ownership records of jewelry and cars
- Dress codes for students, probationers, and parolees
- Forfeitures
- Fines

Labor

- Job programs
- Sanctions for lookouts
- Penalties for using minors
- "Working up the chain" investigations
- Prosecution of minor gang-related crimes
- Field interrogation
- Boys' clubs and athletic leagues

Sellers' Impunity

- "Buy and bust"
- Observation arrests
- Citizen hotlines
- Searches
- Special prosecution policies
- Forfeitures
- Evictions
- More jail/prison capacity
- Special penalties for armed dealers
- More capacity for non-prison sanctions
- Non-criminal punishments
- Finding and using fingerprints
- Electronic surveillance
- Beat cops
- Work with citizens' groups
- Questioning suspected market participants
- Mandatory abstinence and urine monitoring
- Special penalties for dealing near a school

Operating Scope

In addition to threatening dealers with arrest, retail-level enforcement activities can force dealers into more cumbersome operating styles (e.g., they may need to break up the seller's job among runners, holders, and money-handlers). Cutting brush, adding outdoor lights to dark areas, and towing abandoned cars can help eliminate good sites for drug caches, thus complicating outdoor selling.

Drug Supply

A drug dealer needs drugs to sell. The announced goal of much high-level drug enforcement activity is to make those drugs unavailable, or at least in sufficiently unreliable supply at the wholesale level to interfere with retail operations. The list of tactics here is familiar: crop eradication, source-country law enforcement, enforcement directed at the physical transport of drugs from source countries into the United States (interdiction), and the whole panoply of high-level investigative techniques: long-term undercover operations, historical investigation of drug conspiracies, electronic surveillance, financial investigation, and so on. The hope of enforcement agencies is that if the high-level supplier--"Mr. Big"--can be put out of business and his organization disrupted, current retail dealers will find themselves out of stock, and the recruitment of new retail dealers will be slowed. Forfeitures of high-level dealers' assets are designed to decrease the incentive for new Mr. Bigs to replace the old ones.

Retail-level "buy and bust" operations, in which police impersonate drug users for the purpose of catching retail dealers in the act of distribution, attack sellers' impunity. But if, as often happens, some of the retail sellers caught in the trap are offered lenient treatment if they, in turn, help make cases against their suppliers, the result is increased distrust of retail dealers by higher-level dealers. This process of "working up the chain" makes it more difficult for a new retailer, or an established retailer known to be facing charges, to find a "connection" (supplier).

The extent to which drugs can be made scarce, particularly mass-market drugs with established retail networks, has been a matter of debate. In general, the drug enforcement professionals have more optimistic about this approach than have the economists. Several arguments are made for pessimism: the very large financial rewards to successful drug wholesaling, its modest skill requirements, the extent to which a successful smaller sale can finance a subsequent larger purchase, and the apparent flexibility of distribution networks.

Incentive

The income from dealing drugs is either spent or saved. Thus interfering with dealers' ability to spend or save can reduce their willingness to accept the risks of dealing.

Forfeitures and fines are ways of attacking dealers' accumulated wealth. These attacks are more frequent at the wholesale than at the retail level, in part because forfeiture can be laborious for the prosecutor involved and because wholesalers are more likely than retailers to have wealth in quantity and in a form (cash or bank accounts) easy to store pending resolution of the case. (One of the advantages of forfeiture, from the viewpoint of local law enforcement agencies, is that many state laws allow the proceeds to be recycled through their budgets.)

But retail dealers also have wealth, at least in the form of expensive clothing, jewelry, and automobiles. More vigorous efforts to deprive them of these possessions might reduce the value, in their eyes, of their drug-dealing earnings. Some schools have instituted dress codes specifically to reduce the pressures on their students to match drug dealers' clothing expenditures. A limit on the value of clothing and jewelry worn could, in principle, be made part of probation and parole orders, or imposed on juvenile offenders by juvenile courts, but so far as we can determine no such efforts are now underway.

Instead of working from a conviction toward the seizure of wealth, enforcement agencies could use displayed wealth to help identify drug dealers and to provide evidence of their activity. Again, these tactics are more familiar at the wholesale level than at the retail level, but there might be value in forcing dealers to suppress their ostentatious wealth display.

Labor

Retail drug dealing, like other forms of retailing, takes time, time spent waiting for customers. In addition to their own labor, some dealers employ the labor of others, as runners, spotters, steerers, money-handlers, etc.

To the unemployed and bored, their own leisure time may not be very valuable. Either employment opportunity or recreational opportunity can help increase the perceived value of time and thus the reservation wage (the lowest wage one will accept) of potential suppliers of drug-selling labor. Thus the argument for job programs and recreational programs as anti-drug measures.

A recent study showing that many drug dealers also hold legitimate jobs is, however, somewhat discouraging on this score (Reuter 1990). That same study suggests that decreasing the supply of sellers' labor by putting some of them in prison--the incapacitation effect--is also likely to be ineffective, since those now working part-time as dealers constitute an "industrial reserve army for the drug trade."

Sellers' Impunity

The obvious way to decrease sellers' impunity is to increase their probability of arrest by increasing police presence. But in many jurisdictions, current levels of drug arrests have already swamped court and corrections systems; thus, arrests may not be the scarce factor of production in generating deterrence. (See Press 1987)

Punishing drug dealers more severely will involve some combination of three steps: increasing prison populations, increasing the proportion of prisoners serving time for drug dealing (rather than predatory crime), or increasing the supply of non-prison punishments: home confinement, curfews, "community service," intensive probation, etc. One major problem with expanding such programs under current conditions is that to work they all require the availability of prison cells as back-up sanctions for those offenders who fail to comply (e.g., refuse to pay restitution or skip assigned hours of "community service"). Given the prison shortage, those back-up cells may not be available (Cavanagh and Kleiman 1990).

The move toward stiff mandatory minimum sentences for drug dealing, though designed to decrease sellers' impunity, may instead increase the impunity of those selling small quantities or with short criminal histories by using a large share of the available cell-years on a relatively few major dealers.

Although applying the factor of production metaphor to drug markets and efforts to reduce their harmful effects provides a more powerful lens through which to examine the problem, further research is needed to specify the relationships among factors of production, styles of drug dealing, different drug markets, and various enforcement strategies. First, empirical research in different drug markets could identify varying styles of drug dealing by the intensity of their factors of production. Second, research should seek to explain why different styles of dealing emerge in different places at different times on the basis of the intensity of the factors of production. Finally, research that closely consider how enforcement tactics affect the market by pressing on different factors of production would be most useful.

REFERENCES

Anglin, M. D., and W. H. McGlothlin. "Long-Term Follow Up of Clients of High and Low-Dose Methadone Programs." *Archives of General Psychiatry* 38 (1981): 1055-1063.

Anglin, M. D., and G. Speckar . "Narcotics Use, Property Crime, and Dealing: Structural Dynamics Across the Addiction Career." *Journal of Quantitative Criminology* 2 (4) (1986): 355-375.

Anglin, M. D., and G. Speckart. "Narcotics Use and Crime: A Multisample Multimethod Analysis." *Criminology* 26 (2) (1988): 197-233.

BOTEC Analysis Corporation. *Drug Abuse in Jackson County, Missouri: Problem Assessment and Recommendations*. Cambridge, MA.: 1990.

Brown, G. F., and L. P. Silverman. "The Retail Price of Heroin: Estimation and Application." *Journal of the American Statistical Association* 347 (69) (1974): 595-606.

Burns, P., and R. Conner. *Winnable War: A Community Guide to Eradicating Street Drug Market*. Washington, DC.: American Alliance for Rights and Responsibilities, 1991.

Caulkins, J. P. "Distribution and Consumption of Illicit Drugs: Some Mathematical Models and Their Policy Implications." Ph.D. dissertation in Operations Research, Massachusetts Institute of Technology, 1990.

Cavanagh, D. P., and Mark A. R. Kleiman. "A Cost Benefit Analysis of Prison Cell Construction and Alternative Sanctions." Prepared for the National Institute of Justice. Cambridge, MA.: BOTEC Analysis Corporation, 1990.

Czepiel, J.A. et al., ed. *The Service Encounter*. Lexington Books, 1985.

Chaiken, M. R., ed. "Street-Level Drug Enforcement: Examining the Issues." *Issues and Practices*. Washington, D.C.: National Institute of Justice, 1988.

Currie, E. *Confronting Crime: An American Challenge*. New York: Pantheon Books, 1985.

Evaluation and Training Institute (E.T.I.). DARE Longitudinal Evaluation: Annual Report, 1987-88.

Heilbroner, R. L. *The Worldly Philosophers*. New York: Simon and Schuster, 1953.

Kleiman, M. A. R. *Marijuana: Costs of Abuse, Costs of Control*. New York: Greenwood Press, 1989.

Kleiman, M. A. R., M. E. Lawrence, and A. Saiger. "A Drug Enforcement Program for Santa Cruz County." Working Paper 88-01-13, Program in Criminal Justice Policy and Management, John F. Kennedy School of Government, Harvard University, 1987.

Kleiman, M. A. R., and K. D. Smith. "State and Local Drug Enforcement: In Search of a Strategy." In *Drugs and Crime*, edited by M. Tonry and J. Q. Wilson, 69-108, Vol. 13 of *Crime and Justice: A Review of Research*, edited by M. Tonry and N. Morris. Chicago: University of Chicago Press, 1990.

Majority Staffs of the Senate Judiciary Committee and the International Narcotics Control Caucus. *Fighting drug Abuse: A National Strategy*. Draft. 1990.

Moore, M. H. "Achieving Discrimination on the Effective Price of Heroin." *American Economic Review* 63 (2) (1973): 270-277.

Moore, M. H. *Buy and Bust: The Effective Regulation of an Illicit Market in Heroin*. Lexington, MA.: D.C. Heath and Co., 1976.

Mark M. H. "An Analytic View of Drug Control Policies." Paper presented to the Committee on Problems of Drug Dependence, Baltimore, MD, June 4, 1978.

Mark H. M. "Limiting Supplies of Drugs to Illicit Markets in the United States." *Journal of Drug Issues* 9 (1979): 291-308.

Office of National Drug Control Policy. *National Drug Control Strategy Report*. Executive Office of the President, September 1989.

Pentz, M. A., J. H. Dwyer, D. P. MacKinnon et al. "A Multicommunity Trial for Primary Prevention of Adolescent Drug Abuse: Effects on Drug Use Prevalence." *Journal of the American Medical Association* 26 (22) (1989): 3259-3266.

Press, A. *Piecing Together New York's Criminal Justice System: The Response to Crack*. New York: New York Bar Association, 1987.

Reuter, P. "Money From Crime." RAND Corporation Report, Santa Monica, CA.: July 1990.

Reuter, P., and M. A. R. Kleiman. "Risks and Prices: An Economic Analysis of Drug Enforcement." In *Crime and Justice: An Annual Review of Research*, Vol. 7, edited by M. Tonry and N. Morris, 289-340. Chicago: University of Chicago Press, 1986.

Ricardo, D. *Works of David Ricardo*, edited by P. Sraffa. London: Cambridge University Press, 1951.

Spence, A. M. *A Note on the Effects of Pressure in the Heroin Market*. Discussion Paper 588. Cambridge, MA.: Harvard Institute of Economic Research, November 1977.

Speckart, G., and M. D. Anglin. "Narcotics and Crime: A Causal Modeling Approach." *Journal of Quantitative Criminology* 2 (1986): 3-28.

Toborg, M. A., and M. P. Kirby. "Drug Use and Pretrial Crime in the District of Columbia." *Research in Brief*. Washington, D.C.: National Institute of Justice, 1984.

Wilson, J. Q. Introduction. In *Thinking about Crime*. Rev. ed. New York: Basic Books, 1983.

HOW CHANGES IN THE IMPORT PRICE OF ILLICIT DRUGS AFFECT THEIR RETAIL PRICES

Jonathan P. Caulkins

Chapter 3: How Changes in the Import Price of Illicit Drugs Affect Their Retail Prices

3.0 Introduction

This chapter considers how changes in the import price of an illicit drug affect its retail price. Traditionally, interdiction and high-level enforcement were seen as ways of limiting consumption directly by removing drugs and indirectly by incarcerating the dealers that supply them. These views have largely been rejected, however.[1] The markets for the major drugs such as heroin, cocaine, and marijuana are so large and operate so smoothly that even huge seizures such as the 20-ton cocaine seizure in California in the Fall of 1989 do not create noticeable spot shortages.[2] Also, surveys of high school students suggest that availability is not the prime determinant of the prevalence of use.[3]

An alternative theory is that interdiction and high-level enforcement are like taxes. Most heroin, cocaine, and marijuana consumed in the U.S. reaches users through multilayer distribution networks.[4] Enforcement near the source of the network increases dealers' costs and hence increases prices at those levels.[5] Presumably these price increases are passed along in some manner to the consumers. Since, contrary to popular belief, demand for drugs is probably not perfectly inelastic,[6] this in turn reduces consumption.

According to this view, then, the efficacy of border interdiction (and high-level domestic enforcement) depends on two factors: first, how much it increases the import (wholesale) price and second, how much retail prices rise in response to this increase.[7] The first issue

[1] See Reuter, Crawford, and Cave (1988, p.10) and Kleiman (1989, pp.52-55).

[2] International Drug Report, 1989a, p.17.

[3] U.S. Department of Health and Human Services, 1988c.

[4] Descriptions of layered distribution networks go back at least as early as Preble and Casey's (1969) work.

[5] This theory was developed by Reuter and Kleiman (1986).

[6] Reuter and Kleiman (1986, pp.298-301) and Reuter, Crawford, and Cave (1988, pp.20-23) discuss the price elasticities of demand for heroin, cocaine, and marijuana.

[7] Reuter, Crawford, and Cave (1988) consider a third possibility, that enforcement increases variability in the availability of drugs, thereby making them less attractive to use. This possibility is not considered here.

has received considerable attention.[8] In contrast, to the best of the author's knowledge, only two previous studies have formally considered how changes in import prices affect retail prices.[9] This chapter seeks to add to that small literature.

The view put forth by the previous studies (a modified form of what will be called the "additive model") is that price increases are passed along (more or less) dollar for dollar. That is, if import prices rise by $1/unit, retail prices will also rise by about $1/unit. Another view (called the "multiplicative model" below) is that the percentage change in price will be the same at each level. For example, a 10% increase at the import-level will lead to a 10% increase at the retail level. Since retail prices of cocaine and heroin are much greater than their import prices, these views have vastly different implications for the efficacy of interdiction and high-level enforcement.

To illustrate this, suppose the import and retail prices are X and 10X, respectively, and the government is considering an interdiction program that will drive the import price up to 2X. Will the program significantly reduce consumption? According to the additive model, when the import price rises from X to 2X the retail price will rise from 10X to 11X -- a 10% increase which probably would not reduce consumption appreciably. But according to the multiplicative model, when import prices rise from X to 2X, retail prices will double from 10X to 20X, which may noticeably reduce consumption. Hence determining the extent to which the first model, the second, or some blend of the two reasonably reflects reality is quite important.

If drug markets were perfectly competitive, one would expect the additive model to hold. While drug markets are competitive in many respects[10] (for instance they are generally <u>not</u> monopolistic), they fall short of Adam Smith's ideal in several respects. For one, they are characterized by great uncertainty, which suggests that probabilistic analysis may be an appropriate tool for investigating their behavior.

This chapter looks at some simple (decision analytic) lotteries dealers face when they decide whether or not to deal and at what

[8] For example, by U.S. General Accounting Office (1983), U.S. General Accounting Office (1985), Office of Technology Assessment (1986), and Reuter, Crawford, and Cave (1988). Reuter, Crawford, and Cave also mention T. Mitchell and R. Bell's *Drug Interdiction Operations by the Coast Guard* (Center for Naval Analyses, 1980) and a Systems Research Corporation study entitled *Review of Customs Service Marine Interdiction Program* (1985).

[9] Reuter and Kleiman (1986) and Reuter, Crawford, and Cave (1988).

[10] Reuter (1983) makes this point.

price. One can postulate two different sets of assumptions about how dealers perceive the likelihoods and costs of various outcomes of contemplated transactions. One set leads to the additive model; the other to a variant of the multiplicative model called the value-preserving model. The reasonableness of each set of assumptions is discussed, and it will be argued that the dealers' actual behavior may fall between the two sets of assumptions. This suggests that retail price responsiveness may also fall between that predicted by the additive model and that predicted by the value-preserving model.

Then a compromise view, called the multiplicative model, is proposed. The multiplicative model's predictions fall in between those of the additive and value-preserving models, although they are closer to those of the value-preserving model.

The empirical evidence about cocaine prices supports the multiplicative model. No stronger statement can be made, however, for several reasons. First, controlled experiments are not possible. Second, there is essentially no import-level data. Instead wholesale and retail data are compared. Third, the marijuana price trends are less conclusive than the cocaine data, and the data for heroin and other drugs are inadequate for testing the models.

Adjusting for changes in purity and inflation, retail and wholesale cocaine prices moved almost in lock step between 1982 and 1989. Retail prices were consistently about 3.5-5.0 times higher than wholesale prices, even though both prices changed significantly over that period, declining to about one-third of their original values.

This is consistent with the multiplicative model but not the additive model. It does not, however, prove that the multiplicative model is valid for the reasons listed above and because there are other explanations for the proportional relationship between retail and wholesale prices. Specifically, if costs increased by the same fraction at each level of the distribution network, then one might observe such trends in prices.

The next section describes the model used in the two previous studies. The following section examines the problem from a decision analytic framework. This viewpoint leads to two different models depending on what assumptions one makes about the way a dealer's perceptions of certain risks and consequences are affected by a change in the drug's supply price. The additive model, described in Section 3.3, is similar to the one used in the two previous studies. The value-preserving model, described in Section 3.4, is quite different. Section 3.5 discusses the validity of the assumptions underlying the two models. Section 3.6 introduces an intermediate model, called the multiplicative model.

Section 3.7 derives the models' predictions about the relationship between changes in the import price and changes in the retail price. Section 3.8 summarizes the results of the derivations. Section 3.9 describes the empirical evidence on the historical relation between wholesale and retail prices. The last section offers some concluding comments.

3.1 The Model Used in Previous Studies

One model of how changes in price at one level of the distribution network affect prices at subsequent levels (called the wholesale and retail levels, respectively) assumes the retailer simply charges enough more than the wholesale price to cover the costs of dealing, where costs include profits and compensation for risks incurred. More formally, it assumes the retail supply curve is simply the wholesale supply curve shifted upward by a constant representing the cost/unit incurred between purchase and resale. Furthermore, with the exception of the opportunity cost of capital, this cost is assumed to be independent of the wholesale price.

The opportunity cost of capital is the value of earnings foregone because the dealer's money is tied up in the inventory of drugs. It increases with price because at higher prices more capital is tied up during the transaction. Specifically, if r is the annual cost of capital[11] and T years elapse between purchase and resale, then retail prices will increase by $\kappa = (1 + rT)$ times the increase in the wholesale price. Both of the previous studies on this subject assumed that r is between 50 and 100 percent per year and T is 3 months.

As the authors of the previous studies note, these are probably generous upper bounds. Dealers may try to sell drugs as soon as they get them, sometimes even lining up customers before a shipment arrives.[12] So the elapsed time may be less than three months.

The cost of capital is assumed to be high because it is believed that dealers have trouble borrowing from outside lenders. However, many dealers are not cash constrained so they would not need to borrow. In fact, they may have a surplus of cash that they cannot

[11] The annual cost of capital, sometimes called the rental cost of capital, is the cost of using a unit of capital in the same sense that the real wage measures the cost of using a unit of labor. It is commonly identified with the interest rate at which firms can borrow.

[12] This is the impression one gets from Adler (1985) and Mills (1986).

deposit easily because of currency transaction reporting requirements. So not only might their cost of capital be closer to the 10-15% that is usual for a licit enterprise, it might even be lower if the money would otherwise be sitting in a suitcase rather than collecting interest in a bank account.

The view that, except for the cost of holding inventory, costs are passed along dollar for dollar is appropriate for licit goods. To see this, consider another small consumer good supplied primarily from overseas: digital watches. Suppose you are a digital watch dealer. You normally buy boxes of watches off the boat in Los Angeles for $2 per watch and resell them for $3 each. Furthermore, assume that there are many people doing the same thing and that $2 and $3 are the competitive, equilibrium prices.

Now consider what would happen if the price charged at the beach increased to $4 per watch. If you continued selling watches for $3 you would lose money. Even if you increased your price to $4.50 you would probably still lose money, because presumably the previous $1 price differential was required to cover your costs and normal profit.

On the other hand, if you tried to increase your prices above $5 plus the increase in inventory holding costs, your competitors could undercut your price.

In a competitive market, when the import price increases, watch dealers would increase the retail price just enough to cover their additional costs. If their other costs of doing business, such as the costs of labor, advertising, and distribution do not depend on the price of the watches, then the only costs that go up are the direct purchase cost and the cost of holding inventory.

Hence this view suggests that the change in retail price (ΔP_R) equals the change in wholesale price (ΔP_W) after adjusting for the increase in holding costs. This implies that the new retail price (P_R') is

$$P_R' = P_R + \Delta P_R = P_R + (1 + rT)\,\Delta P_W, \qquad (3.1)$$

where (1 + rT) is a positive constant, typically a little larger than one. The value of (1 + rT) = 1.125 used by Reuter and Kleiman[13] considers only the opportunity cost of capital. There are other post-import effects of an increase in the import price, however. For example, the risk of being robbed or defrauded increases with the value of the

[13] Reuter and Kleiman, 1986, p.305.

drugs. Reuter, Crawford, and Cave[14] try to account for them by rounding (1 + rT) up to 2.0.

3.2 A Decision Analytic Viewpoint

Decision analysis is a technique for analyzing decisions that explicitly considers risk and uncertainty.[15] At some level, risk and uncertainty are present for everyone, every day. For participants in illicit drug markets, however, the costs of uncertain but not uncommon events such as being arrested or being murdered by another dealer far outweigh the day-to-day costs of buying adulterants and transporting the drugs. Hence, decision analysis may be an appropriate tool for trying to understand the behavior of drug markets.

One caveat is in order. Decision analysis is a prescriptive not a descriptive technique. That is, it tries to answer the question, "How should one make a decision?" not "How are decisions actually made?" However, to the extent that people rationally act in their own self-interest, they often behave in accordance with the tenets of decision analysis.

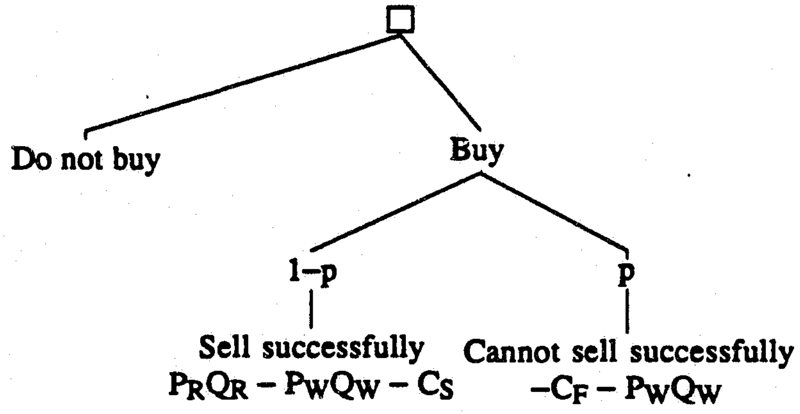
Figure 3.1 proposes a decision tree for a dealer who is deciding whether to buy and resell drugs. For simplicity it is assumed that the dealer can estimate beforehand how much the supplier wants to sell[16] (Q_W), the final price the two will negotiate (P_W), and the average price (P_R) the dealer will receive for the amount (Q_R) the dealer decides to resell.

Frequently dealers dilute ("step on") the drugs with adulterants. To avoid confusion, quantities should be understood to refer to the weight of the drugs themselves (excluding adulterants) and all prices are the price per pure unit weight of the drugs. Hence $Q_R \leq Q_W$, and Q_R would only be less than Q_W if the dealer used some of the drugs or there is some leakage or waste in the course of a successful deal.

[14] Reuter, Crawford, and Cave, 1988, p.19.

[15] It is assumed that the reader is familiar with elementary decision theory. If not, Raiffa (1968) and Keeney and Raiffa (1976) offer authoritative introductions to the subject.

[16] The assumption that the dealer will buy quantity Q_W or nothing at all is a simplification; generally other quantities would be available. However, the argument below simply derives price-quantity pairs the dealer would be willing to buy and sell. Since one of the axioms of decision analysis is that adding new alternatives never inverts established preferences, omitting options to purchase other quantities does not invalidate the conclusions.



P_W = price dealer pays supplier (wholesaler) for drugs
Q_W = quantity of drugs supplier offers at the price P_W
P_R = average price at which dealer sells to customers (retail price)
Q_R = amount dealer sells if he or she sells successfully
p = probability dealer fails to sell successfully
C_F = costs, other than direct purchase costs, incurred when dealer fails to sell successfully
C_S = costs, other than direct purchase costs, incurred when dealer sells successfully

Figure 3.1: Decision Tree Faced by a Dealer

The branch labelled "cannot sell successfully" represents all the outcomes that are unfavorable to the dealer. These include being imprisoned for various lengths of time, arrested and put on probation, arrested and released, robbed or defrauded by the supplier (e.g., drugs purchased are of lower quality than the supplier claimed or the supplier takes the dealer's money without delivering the drugs), robbed or defrauded by a buyer (e.g., buyer steals drugs or buys them on credit and cannot make payments), having the dealer's cache of drugs stolen, etc. Thus, p is the probability that something goes wrong for the dealer, and C_F is the expected cost, beyond the purchase cost, incurred by the dealer if something goes wrong. Note, C_F is not simply a dollar cost because it includes the disutility of a variety of unfavorable outcomes.

As Figure 3.1 shows, the decision maker has the option of buying drugs or not. If the decision maker chooses to buy drugs, there is a probability p that the decision maker cannot sell them successfully and receives the (negative) reward $-C_F - P_W Q_W$. Likewise, with probability (1 − p) the decision maker is able to sell them "successfully" and receives reward $P_R Q_R - P_W Q_W - C_S$. Thus if the decision maker would willingly accept the chance to play a lottery that paid $P_RQ_R - P_WQ_W - C_S$ with probability $(1 - p)$ and $-C_F - P_WQ_W$ with probability p, then that decision maker would presumably choose the "buy" branch and become a dealer.

Of course it would be extremely difficult to learn enough about any given active dealer's preferences and risk perceptions to explicitly model the subtree represented by the "cannot sell successfully" branch. Moreover, determining the requisite preferences and risk perceptions for all dealers is out of the question. Nevertheless, the next two sections suggest that some conclusions can be drawn about a dealer's response to a change in the supply price as long as the dealer's preferences and perceptions of risk about events in the subtree do not change.

3.3 The Additive Model

Consider a person who when confronted with the choice depicted in Figure 3.1 decides to deal drugs. One can infer that the person prefers the "buy" branch to the "do not buy" branch. Now suppose the supply price changes to $P'_W = P_W + \Delta P_W$, making the "buy" branch less attractive. For what quantities (Q'_W and Q'_R) and retail price (P'_R) would the dealer still prefer the "buy" branch to the "do not buy" branch despite the higher supply price P'_W?

In general one cannot answer that question without knowing a great deal about the dealer's preferences and perceptions of risk. But suppose that if the dealer buys and sells the same amount, the dealer's perceptions of the likelihood and consequences of the unfavorable outcomes in the subtree represented by the branch "cannot sell successfully" and the costs of a successful deal (except the direct purchase costs and the opportunity cost of the capital tied up in inventory) remain the same after the supply price increases. Then consider how the dealer would respond to the opportunity to buy and sell the same quantity as before ($Q'_W = Q_W$ and $Q'_R = Q_R$) if the average resale price rises enough to compensate for the higher direct purchase cost. Specifically, if

$$P'_R = P_R + \frac{Q_W}{Q_R}\left(\frac{1}{1-p} + rT\right)\Delta P_W. \qquad (3.2)$$

As before the rT term accounts for increased inventory costs. The $({}^{1}/_{(1-p)})$ term is necessary because the dealer only sells successfully, and hence receives the higher price, with probability (1−p). Figure 3.2 shows the new decision tree.

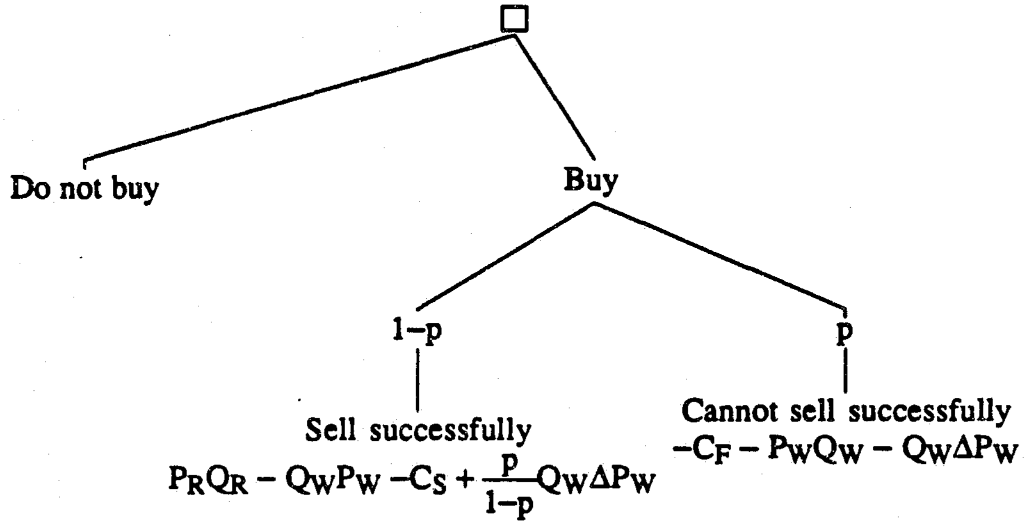

Do not buy
Buy
1−p
p
Sell successfully
$P_R Q_R - Q_W P_W - C_S + \frac{p}{1-p} Q_W \Delta P_W$
Cannot sell successfully
$-C_F - P_W Q_W - Q_W \Delta P_W$

Figure 3.2: Revised Decision Tree With $P_R' = P_R + \frac{Q_W}{Q_R}\left(\frac{1}{1-p} + rT\right)\Delta P_W$

Since the decision maker prefers the "buy" branch in Figure 3.1 to the "do not buy" branch, for two reasons it is likely, although not certain, the dealer will also prefer the "buy" branch in Figure 3.2. The first of these reasons is that the expected value of the two lotteries is the same, so a risk neutral decision maker would value the lotteries equally. Most people are risk averse and so prefer the lottery in Figure 3.1 to the lottery in Figure 3.2. However, the simple fact that the decision maker is a drug dealer suggests that the decision maker is not too risk averse. Second, since $Q_W \Delta P_W$ and $({}^{p}/_{(1-p)})Q_W \Delta P_W$ are likely to be small relative to $-C_F - P_W Q_W$ and $P_R Q_R - P_W Q_W - C_S$, the consequences in the two lotteries are similar, and one can reasonably approximate a risk averse utility function by a risk neutral one if the range of consequences is small.

If either of these reasons hold and the dealer's perceptions of the risks and costs in the "cannot sell successfully" subtree are not affected by an increase in the wholesale price, the dealer would be willing to deal the same quantity as before after increasing the retail price by the amount indicated in Equation 3.2. One can only argue the dealer would be willing to supply the drugs under these conditions, not that they will be supplied. Whether the deal actually takes place also depends on the preferences of the buyers, i.e. on demand. The interaction with demand will be discussed in Section 3.7.

If these assumptions hold at all levels of the domestic distribution network, the analysis can be applied to each level, and the results combined. Then, if the import price increases by ΔP_I, the model predicts that the domestic distribution network would be willing to import and retail the same amounts ($Q_R' = Q_R$) if the retail price increased to

$$P_R' = P_R + \kappa \Delta P_I \qquad (3.3)$$

where

$$\kappa = \frac{Q_I}{Q_R}\left[\prod_{i=1}^{N}\left(\frac{1}{1-p_i} + r_i T_i\right)\right],$$

N = number of levels in the domestic distribution network between import and retail sale,

ΔP_I = change in import price,

Q_I = amount imported,

Q_R = amount sold at the retail level,

p_i = probability dealer at level i fails to sell successfully,

r_i = annual cost of capital for dealer at level i, and

T_i = time between purchase and resale at level i.

Equation 3.3 suggests referring to this model as the "additive model." The additive model is structurally similar to the model used in the previous studies except it relates the retail prices the domestic distribution network would be willing to offer before and after the import price change not the actual equilibrium retail prices.

Also, the constant κ in Equation 3.3 includes several factors that the constant in Equation 3.1 does not. The $\frac{Q_I}{Q_R}$ term in the expression for κ accounts for leakages, both figurative and literal, that occur at various points in the network even if all sales are successful. If one views the network as a black box with money flowing in from the customers and out to the smugglers, the change in retail price must be $\frac{Q_I}{Q_R}$ times the change in import price to preserve the same net flow of money into the black box.

If $p_i = 0$ for all i, then the middle term is $\prod_{i=1}^{N}(1+r_i T_i)$. This reduces to 1 + rT if compounding is ignored.

The $^1/(1 - p_i)$ terms further inflate the price to keep the expected revenues at each level constant. They would fall otherwise because with probability p_i the sale fails and no money is collected. Note, these terms may not be very significant. While the probability of a dealer's being arrested during a year of active dealing may be significantly greater than 0, the p_i's for a single deal are much smaller. On the other hand, the p_i's also include the probability of being robbed or defrauded, which may be significantly greater than the probability of being arrested.

To summarize,. this section used a decision analytic viewpoint to derive a model that is essentially the same as the one used in the previous studies. They differ only in the expressions for the proportionality constant multiplying the change in the import price, and for reasonable parameter values, the expressions represent similar values.

3.4 The Value-Preserving Model

The key assumption in the derivation above was that the dealer's perceptions of the likelihoods of certain outcomes and their costs (other than direct purchase costs and the cost of capital) are unaffected by an increase in the supply price if the quantities purchased and sold remain the same. One would expect this to be true if costs depend primarily on weight or volume as they might for a company that purchases oil in the Middle East, ships it to the U.S., and sells it here. The price/unit weight of drugs is so high, however, that it is at least plausible that the dominant costs will be proportional to the dollar value of the transaction not the quantity transacted. This section argues that if this is indeed the case, a quite different model of how price increases are passed along may be more accurate.

If the supply price increases to $P_W' = P_W + \Delta P_W$, but the dealer buys proportionately less $(Q_W' = \frac{P_W}{P_W'} Q_W)$, the dollar value of the purchase remains constant $(Q'_W P'_W = Q_W P_W)$. Likewise, if the dealer tries to sell the same fraction of the amount purchased $\left(Q'_R = Q'_W\left(\frac{Q_R}{Q_W}\right) = Q_R\left(\frac{P_W}{P'_W}\right)\right)$ and the new retail price is proportionately higher $\left(P'_R = P_R\left(\frac{P'_W}{P_W}\right)\right)$, then revenues from a successful sale remain the same $(P'_R Q'_R = P_R Q_R)$.

Then if the likelihoods and consequences of the unfavorable outcomes and the costs of a successful sale depend only on the dollar value of the transaction, the resulting decision tree (shown in Figure 3.3) is identical to the one in Figure 3.1; their leaves have exactly the same values.

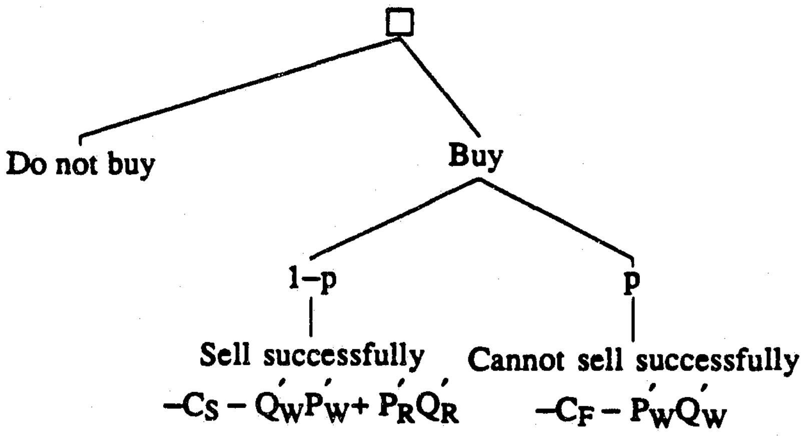


Figure 3.3: Revised Decision Tree for Value-Preserving Model

Hence, if the dealer's perceptions of the likelihoods and consequences in the subtree represented by the branch "cannot sell successfully" depend only on the dollar value of the transaction, the dealer's willingness to deal under the conditions in Figure 3.1 imply a willingness to deal under the conditions depicted in Figure 3.3.

This suggests that if the supply price increases by X%, the dealer would be willing to supply proportionately less at an average price that is X% higher than before. This does not mean that the actual price and quantity will change in this manner; that depends on demand as well as supply. It says only that the dealer would be willing to deal under those circumstances. Section 3.7 will consider demand as well.

By applying the same analysis to each level of the distribution network and concatenating the results, one reaches the same conclusion about the relation between price-quantity pairs at the import and retail levels. Namely, dealers would be willing to offer

$$Q_R' = \frac{Q_I'}{Q_I} Q_R \qquad (3.4)$$

at price

$$P_R' = \frac{P_I'}{P_I} P_R. \qquad (3.5)$$

This model is called the "value-preserving model" for obvious reasons.

Summarizing the results above, if dealers' perceptions of the likelihoods and consequences of the events in the subtree "cannot sell successfully" and the costs of a successful sale depend principally on the dollar value of the transactions, one would expect the value-preserving model to hold. If the value-preserving model holds, the domestic distribution network would be willing to respond to an increase in the import price by offering a proportionately smaller volume at a proportionately higher price.

If, on the other hand, dealers' perceptions of the likelihoods and consequences depend primarily on the quantity transacted not on the dollar value of the transaction, one would expect the additive model to hold. In that case one would expect the domestic distribution network to try to pass along an import price increase (inflated by a constant factor κ) to the users.

Note, it is not important that the dealers actually estimate their risks or even that they identify them as depending primarily on quantity or primarily on dollar value. They only need to understand their risks and costs well enough to run their business. The word perceptions is used above simply because decisions are analyzed from their perspective, not an objective point of view.

The next section will discuss how realistic the two sets of assumptions are.

3.5 Validity of the Two Models' Assumptions

This section considers the validity of the assumptions underlying the additive and value-preserving models. It distinguishes between two kinds of unfavorable outcomes for dealers: those resulting from the actions of authorities and those resulting from the actions of other participants in the drug trade. Roughly speaking, the probabilities and consequences of the first group satisfy the assumptions of the additive model while the probabilities and consequences of the second satisfy those of the multiplicative model. Hence, a blend of the two models may be more accurate than either alone.

Since one cannot speculate intelligently about dealers' perceptions of risks and consequences, it will be assumed that those perceptions reflect reality sufficiently that if the true probabilities and consequences depend predominantly on quantity or dollar value, so will the dealers' perceptions.

There are five broad categories of costs and unfavorable outcomes for dealers: fixed costs, arrest, seizure or forced loss of drugs by authorities, robbery or fraud, and homicide. In addition, some costs are incurred even when the sales are successful.

Fixed Costs

Dealers' costs that do not vary with quantity or value satisfy both models assumptions. The additive model does not require that the risks and consequences be proportional to quantity. It assumes only that they do not change if prices increase but quantity remains the same. This is clearly the case for fixed costs. For similar reasons, fixed costs satisfy the assumptions of the value-preserving model.

Arrests

The consequences of arrest do not depend on the value of the drugs. The consequences do not always vary greatly with quantity either, but in as much as they do, they increase with quantity because the maximum punishment for convicted drug offenders increases with quantity in a staircase fashion. Only a small fraction of those arrested actually serve the full sentence, but the possible sentence influences the actual sentence, plea bargaining, bail requirements, and so on. Likewise, enforcement agents' incentive systems generally depend on quantity, so arresting officers and agents are likely to work harder to make strong cases and see them through if they involve larger quantities of drugs.

The probability of arrest also depends more on quantity than dollar value because it depends heavily on the number of connections the dealer must make and maintain. The more sales the dealer makes, the more likely it is that one of the customers will be an undercover agent, a customer will be arrested and "turn" or choose to become an informant,[17] and that the dealer will be arrested as a result of direct observation by uniformed or undercover officers.

Hence the likelihoods and costs depend more on quantity than value and thus probably come closer to satisfying the assumptions of the additive model.

Seizure or Forced Loss of Drugs

Dealers can lose their drugs as a result of enforcement efforts that do not result in arrest (e.g. the dealer's employee is arrested with the drugs or the dealer is forced to abandon the drugs). For the reasons mentioned above, the likelihood of this probably satisfies the additive model's assumptions, but the consequence depends directly on the dollar value and so satisfies the value-preserving model's assumptions. Thus when the supplier's price increases, the adjustments described with the additive model do not fully compensate the dealer, and the adjustments described with the value-preserving model are overly generous. Hence in some sense, the likelihood and consequences of these events fall "in between" the two models' assumptions.

Robbery and Fraud

The probabilities and consequences of being robbed or defrauded increase with price, but that increase may be essentially cancelled by a decrease in quantity that preserves the dollar value of the transaction. This is clearest for the consequence of having one's cache stolen. The consequence depends entirely on the dollar value of the drugs. The likelihood of having one's cache stolen also increases with price because the temptation to burglars increases, so it does not satisfy the assumptions of the additive model. On the other hand, the likelihood decreases with quantity since it is easier to conceal a smaller amount. This decrease may not exactly offset the price related increase, but this likelihood comes closer to satisfying the assumptions of the value-preserving model than those of the additive model.

[17] The reward offered to an informant and thus the incentive to inform may depend somewhat on the dollar value of the transactions and thus on price, but that is probably a second order consideration.

The consequence of other forms of robbery and fraud (such as suppliers' selling substandard quantity, buyers stealing drugs or never paying, etc...) also depend on the dollar value not just the quantity. The temptation to rob or defraud increases with price. However, as overall quantities decrease dealers can be more selective in their choice of suppliers and buyers and thus moderate the effects of the increased temptation. Again, it is not clear that these effects will exactly offset, but the net effect will come closer to satisfying the assumptions of the value-preserving model than the additive one.

Homicide

The discussion of the likelihood of robbery or fraud probably extends to the likelihood of being murdered.[18] The consequence to the dealer of being murdered certainly does not change if price increases, so the likelihood and consequence of being murdered probably satisfies the value-preserving models' assumptions.

Costs Incurred During Successful Sales

The discussion above focused on the probability and cost of failing to sell successfully. The models also assume that the cost of selling successfully (the dealers' regular operating expenses) are invariant. These costs are likely to be smaller than the risks of arrest, robbery, and fraud, so they will have less impact on the accuracy of the two models, but they are worth discussing.

Obviously the physical cost of moving and concealing the drugs depends only on quantity, and hence satisfies the assumptions of the additive model. However, these costs are quite small, except perhaps at the highest levels. A more significant cost that satisfies the assumptions of the additive model is the cost of the dealer's time. The amount of time required to sell a shipment of drugs probably depends more directly on the quantity (number of sales) than the price.

Upper-level dealers have employees who must be paid. Employees' wages must be high enough to compensate the workers for their time, the risks they incur, and their loyalty.

When the labor needed is proportional to the quantity of drugs (as it is for jobs like packaging), that component of wages satisfies

[18] Homicide can be a significant risk. Reuter et al. (1990) roughly estimate that a typical cocaine retailer in Washington D.C. receives compensation of $10,500/yr. for the risk of homicide, $2,100/yr. for the risk of injury, and $7,000/yr. for the risk of imprisonment.

the assumptions of the additive model. Some jobs, such as guarding a cache, require about the same amount of labor for large ranges of quantities and values. For other jobs, such as those of body guards and collection agents, the labor requirement probably depends more on value than quantity. To see this, suppose drugs were very inexpensive. Then few users would default so there would be less work for collection agents. Also, there would be less incentive to murder the dealer to take over the dealer's business, so there would be less work for bodyguards.

The risks to which workers are exposed are similar to those experienced by the dealer. Hence risks from enforcement probably depend more on the quantity while risks from other participants in the market depend more on value.

Finally, payments required to keep employees from absconding with drugs or money depend on the value of the transactions.

Thus the extent to which wages fit either of the two models' assumptions mirrors that of the dealer's costs as a whole.

Other costs, such as the cost of financing the dealer's own habit (assuming the dealer consumes a constant proportion of the amount dealt) depend on the dollar value of the transaction, and so satisfy the assumptions of the value-preserving model. For retail dealers this can be a significant fraction of total costs.

The costs of avoiding robbery, fraud, and arrest are arguably the most significant costs of a successful transaction. As mentioned above, the costs of concealment depend on quantity, but most of the other avoidance costs such as ensuring employee loyalty (by direct payment or by maintaining a capacity for violence) and bribes probably depend more on the dollar value of the transaction than quantity or price alone.

Thus the risks and consequences of a successful transaction do not neatly satisfy either set of assumptions. However, they probably play a smaller role in the dealer's decision than the risks and consequences of not selling successfully. Also, they contain many components that are relatively insensitive to changes in price or quantity. So it is probably not reasonable to argue against either the additive or the value-preserving view on the grounds that the costs of a successful transaction do not satisfy the requisite assumptions.

To summarize, generally speaking, the probabilities and consequences of actions taken by authorities satisfy the assumptions of the additive model while those resulting from the actions of other participants in the drug trade (robbery, fraud, homicide) come closest to satisfying the assumptions of the value-preserving model. This suggests that the true relationship between retail and import prices will be a blend of the two model's predictions, perhaps weighted toward the value-preserving model because the likelihoods and consequences of being robbed or defrauded are generally thought to be greater than those of arrest.[19]

3.6 An Intermediate (Multiplicative) Model

It has been argued that neither the additive nor the value-preserving model's assumptions hold completely. That is, it is neither true that all probabilities and costs depend only on the quantity nor that they all depend only on the value of the transactions. To the extent that some probabilities and costs depend on the value of the transaction, the additive model understates the impact of a price change, and to the extent that some probabilities and costs depend on the quantity, the value-preserving model overstates the impact of a price change.

One intermediate model of how prices are passed along is

$$P_R' = \frac{P_I'}{P_I} P_R \quad \text{and} \tag{3.6}$$

$$Q_R' = Q_R. \tag{3.7}$$

It will be called the multiplicative model because it suggests that prices are passed along on a percentage basis. The multiplicative model leads to greater shifts in the retail supply curve than the additive model because

$$P_R' = P_R' = \frac{P_I'}{P_I} P_R = \frac{P_I + \Delta P_I}{P_I} P_R = P_R + \frac{P_R}{P_I} \Delta P_I. \tag{3.8}$$

Retail prices are much higher than import prices, so the coefficient of ΔP_I in this expression is larger than κ, the corresponding coefficient for the additive model.

On the other hand, the intermediate model leads to smaller shifts in the retail supply curve than the value-preserving model. Suppose P_I increased, so $P_I' > P_I$. Then the retail price offered increases by the same amount for both the multiplicative and the

[19] Garreau, 1989.

value-preserving models, but according to the value-preserving model, the quantity offered will decrease as well. The multiplicative model is less extreme. It predicts that the quantity offered will remain the same as long as the price increases by the specified amount.

There is no "story" to justify this intermediate model. If some probabilities and costs depend on quantity and others depend on the value of the transactions, then one cannot construct a simple lottery that keeps the "cannot sell successfully" branch constant. As will be seen, however, the multiplicative model is analytically convenient and matches the empirical data well.

3.7 Derivation of the New Retail Equilibrium

The (P_I, Q_I) and (P_R, Q_R) combinations described above are price-quantity pairs at which the dealer is willing to buy and sell, respectively. (P_I', Q_I') and (P_R', Q_R') are too. One cannot conclude, however, that if the dealer were originally buying and selling at (P_W, Q_W) and (P_R, Q_R) and conditions changed so the supplier offered (P_W', Q_W'), that the dealer would then sell quantity Q_R' at price P_R'. The dealer would be willing to operate at that price and quantity, but the customers might not be, so (P_R', Q_R') might not be a market equilibrium point. The new equilibrium price and quantity depend on demand as well as supply.

The additive, multiplicative, and value-preserving models describe ways a shift in the import supply curve might affect the retail supply curve. If one assumes particular mathematical forms for the original supply curve, demand curve, and the fraction of drugs imported that are ultimately sold at the retail level, they allow one to derive expressions for the retail price and quantity before and after a shift in the import supply curve. This is done next.

The derivation is somewhat technical, so readers may want to skip directly to Section 3.8 which explains the significance of the results.

A supply curve is the set of price-quantity pairs the market is willing to provide. Ideally one would like to express the retail supply curve as a function of the supply curve at the import level. That is, one would like to model the domestic distribution network (DDN) as a function $f:R^2 \text{--->} R^2$ that maps price-quantity pairs (P_I, Q_I)

smugglers are willing to provide into price-quantity pairs $f = (f_1, f_2) = (P_R, Q_R)$ at the retail level. (See Figure 3.4.)

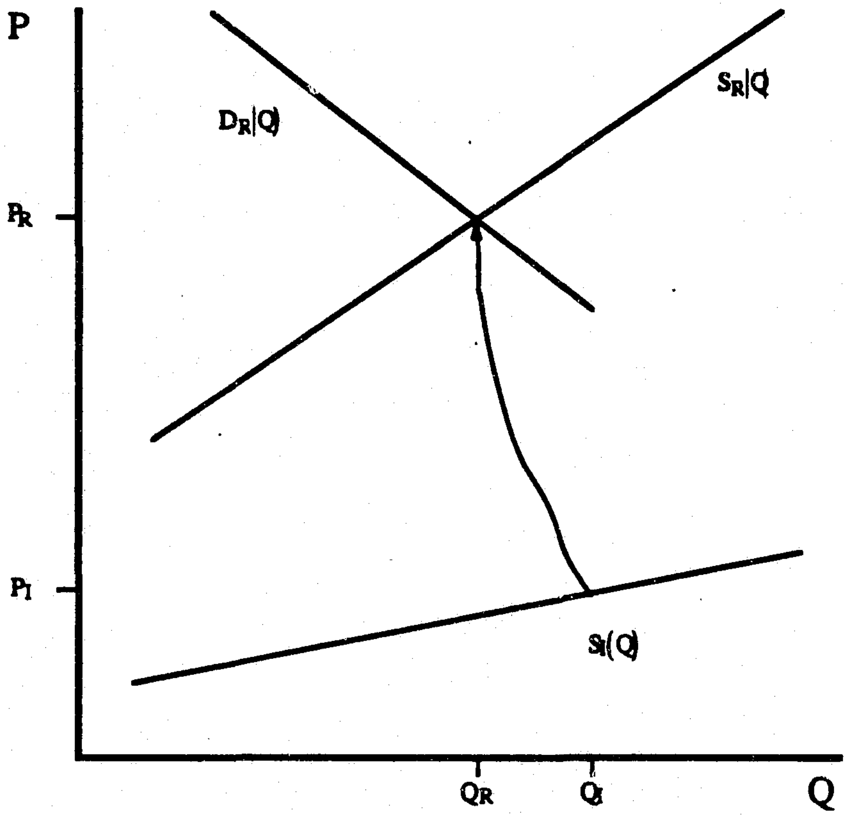


Figure 3.4: Relation Between Import Supply, Retail Supply, and Retail Demand Curves

$S_R(Q)$ is the retail supply curve
$D_R(Q)$ is the retail demand curve
$S_I(Q)$ is the import supply curve
P_R is the current equilibrium retail price
Q_R is the current equilibrium retail quantity
P_I is the current equilibrium import price
Q_I is the current equilibrium quantity imported

Although economists usually express quantity as a function of price, working with the inverse relationship, price expressed as a function of quantity, is equally valid. The latter is used here because it simplifies the algebra.

It is reasonable to postulate some properties of **f**. For example, one would expect f_1 to be increasing in P_I and greater than P_I.

Likewise, f_2 is probably increasing in Q_I and no greater than Q_I. However, it is not practical to find $\mathbf{f}$ explicitly. Fortunately, this is not necessary.

The models proposed above suggest how the retail supply curve changes when the import price changes. Specifically the additive model suggests that

$$\mathbf{f}_1(P_I + \Delta P_I, Q_I) - \mathbf{f}_1(P_I, Q_I) = \kappa \, \Delta P_I, \tag{3.9a}$$

the multiplicative model suggests that

$$\mathbf{f}_1(P_I + \Delta P_I, Q_I) - \mathbf{f}_1(P_I, Q_I) = \frac{\Delta P_I}{P_I} \mathbf{f}_1(P_I, Q_I), \tag{3.9b}$$

and the value-preserving model suggests that

$$\mathbf{f}_1\left((1+\alpha)P_I, \frac{1}{(1+\alpha)}Q_I\right) = (1+\alpha)\,\mathbf{f}_1(P_I, (1+\alpha)Q_I). \tag{3.9b}$$

The value-preserving model leads to complicated and nonintuitive expressions for the new equilibrium retail price and quantity, so the analysis for the value-preserving model is not presented here. It is easy to see, however, that it predicts greater retail price changes than either the additive or multiplicative models do.

Since $\mathbf{f}(P_I, Q_I)$ is just the current retail price-quantity equilibrium pair, Equation 3.9 allows one to estimate how a change in the import supply curve affects the retail supply curve.

There are at least two reasonable conjectures for how increasing interdiction affects the import supply curve. It may shift the curve (and hence the price offered for a given quantity) up by a constant amount for all quantities,

$$S_I'(Q) = S_I(Q) + \Delta P_I \tag{3.10}$$

or by a constant percentage for all quantities,

$$S_I'(Q) = (1+\alpha)S_I(Q). \tag{3.11}$$

These two views will be referred to as the first and second interdiction models, respectively. The next few pages derive expressions for the original and new price-quantity pairs at the retail and import levels for the additive and multiplicative models with both interdiction models.

The derivations assume the import supply, retail supply, and retail demand curves are linear, and the DDN retails a fixed percentage (β) of imports, i.e.

$$S_I(Q) = a_I\, Q + b_I \qquad (3.12a)$$
$$S_R(Q) = a_R\, Q + b_R, \qquad (3.12b)$$
$$D_R(Q) = a_D\, Q + b_D \quad \text{with } a_D \le 0, \text{ and} \qquad (3.12c)$$
$$f_2(P, Q) = \beta\, Q. \qquad (3.12d)$$

Equating (3.12b) and (3.12c) implies that the initial equilibrium retail price and quantity are

$$P_R = \frac{a_R b_D - a_D b_R}{a_R - a_D} \quad \text{and} \qquad (3.13)$$

$$Q_R = \frac{b_D - b_R}{a_R - a_D}. \qquad (3.14)$$

Thus by (3.12a) and (3.12d), the import price and quantity are

$$P_I = \frac{a_I\,(b_D - b_R)}{\beta\,(a_R - a_D)} + b_I \quad \text{and} \qquad (3.15)$$

$$Q_I = \frac{b_D - b_R}{\beta\,(a_R - a_D)}. \qquad (3.16)$$

Under the first interdiction model (Equation 3.10), $a_I' = a_I$ and $b_I' = b_I + \Delta\, P_I$. Under the second interdiction model (Equation 3.11) $a_I' = (1 + \alpha)\, a_I$ and $b_I' = (1 + \alpha) b_I$.

For the additive model, if (3.12d) holds

$$S_R'(Q_R) = S_R(Q_R) + \kappa\left[S_I'\left(\frac{Q_R}{\beta}\right) - S_I\left(\frac{Q_R}{\beta}\right)\right] \qquad (3.17)$$

so $a_R' = a_R$ and $b_R' = b_R + \kappa\,\Delta P_I$ for the first interdiction model and $a_R' = a_R + \kappa\alpha\frac{a_I}{\beta}$ and $b_R' = b_R + \kappa\,\alpha\,b_I$ for the second.

For the multiplicative model,

$$S_R'(Q_R) = \frac{S_I'\left(\frac{Q_R}{\beta}\right)}{S_I\left(\frac{Q_R}{\beta}\right)}\,S_R(Q_R) \qquad (3.18)$$

so for the first interdiction model

$$S_R'\left(Q_R'\right) = \frac{\frac{a_I}{\beta}Q_R' + b_I + \Delta P_I}{\frac{a_I}{\beta}Q_R' + b_I}\left(a_R Q_R' + b_R\right). \qquad (3.19)$$

This is a nonlinear function of Q_R', but if ΔP_I is small enough that $Q_R' \approx Q_R$, then $\frac{a_I}{\beta}Q_R' + b_I \approx P_I$ and

$$S_R'\left(Q_R'\right) \approx \left(1 + \frac{\Delta P_I}{P_I}\right)\left(a_R Q_R' + b_R\right) \qquad (3.20)$$

so $a_R' \approx \left(1 + \frac{\Delta P_I}{P_I}\right)a_R$ and $b_R' \approx \left(1 + \frac{\Delta P_I}{P_I}\right)b_R$. For the second interdiction model $a_R' = (1 + \alpha)\,a_R$ and $b_R' = (1 + \alpha)\,b_R$. Table 3.1 summarizes the changes in the parameter values.

Substituting these new parameter values into Equations (3.13) - (3.16) gives expressions for the new equilibrium price and quantity at both the retail and import level. (See Table 3.2.)

Table 3.1:
Changes in Parameter Values When Import Supply Is Restricted

	Interdiction Model #1	Interdiction Model #2
Import Level		
a_I'	a_I	$(1 + \alpha)\, a_I$
b_I'	$b_I + \Delta\, P_I$	$(1 + \alpha)\, b_I$
Additive Model		
a_R'	a_R	$a_R + \kappa\, \alpha\, \frac{a_I}{\beta}$
b_R'	$b_R + \kappa\, \Delta P_I$	$b_R + \kappa\, \alpha\, b_I$
Multiplicative Model		
a_R'	$\left(1 + \frac{\Delta P_I}{P_I}\right) a_R$*	$(1 + \alpha)\, a_R$
b_R'	$\left(1 + \frac{\Delta P_I}{P_I}\right) b_R$*	$(1 + \alpha)\, b_R$

*Denotes approximation valid for small ΔP_I.

The corresponding expressions in Table 3.2 for the model used in the two previous studies come directly from Equation 3.1:

$$P_R' = P_R + (1 + rT)\Delta\widehat{P}_I. \tag{3.1}$$

If the retail demand curve is linear, then $\Delta Q_R = \frac{\Delta P_R}{a_D}$. This implies that

$$Q_R' = Q_R + \frac{\Delta P_R}{a_D} = Q_R - \frac{(1 + rT)}{|a_D|}\Delta\widehat{P}_I. \tag{3.21}$$

The symbol $\Delta\widehat{P}_I$ is used for the observed change in the import price $P_I' - P_I$. It equals the shift in the import supply curve, denoted by ΔP_I, if and only if the import supply is perfectly elastic or demand is perfectly inelastic. Similarly, $\hat{\alpha}$ denotes the observed percentage change in the import price, in contrast with α, which is the percentage increase in the import supply curve.

Table 3.2: Price and Quantity at Retail and Import Level After the Import Supply Curve Shifts

	Interdiction Model #1	Interdiction Model #2
Model Used in Previous Studies		
P_R'	$P_R + (1+rT)\Delta\widehat{P}_I$	$P_R + (1+rT)\hat{\alpha}\, P_I$
Q_R'	$Q_R - \frac{(1+rT)}{\lvert a_D\rvert}\Delta\widehat{P}_I$	$Q_R - \frac{(1+rT)\hat{\alpha}}{\lvert a_D\rvert} P_I$
P_I'	$P_I + \Delta\widehat{P}_I$	$(1+\hat{\alpha})P_I$
Additive Model		
P_R'	$P_R + \frac{\kappa\lvert a_D\rvert}{a_R - a_D}\Delta P_I$	$P_R + \frac{\kappa\lvert a_D\rvert\alpha}{(a_R - a_D) + \frac{\kappa\alpha}{\beta}a_I} P_I$
Q_R'	$Q_R - \frac{\kappa}{a_R - a_D}\Delta P_I$	$Q_R - \frac{\kappa\alpha}{(a_R - a_D) + \frac{\kappa\alpha}{\beta}a_I} P_I$
P_I'	$P_I + \Delta P_I - \frac{\kappa a_I}{\beta(a_R - a_D)}\Delta P_I$	$(1+\alpha)P_I - \frac{\kappa a_I \alpha}{\beta(a_R - a_D) + \kappa\alpha a_I} P_I$
Q_I'	$Q_I - \frac{\kappa}{\beta(a_R - a_D)}\Delta P_I$	$Q_I - \frac{\kappa\alpha}{\beta(a_R - a_D) + \kappa\alpha a_I} P_I$

Table 3.2: (cont.)

Multiplicative Model

P_R'	$P_R + \dfrac{\lvert a_D \rvert \frac{P_R}{P_I}}{\left(1+\frac{\Delta P_I}{P_I}\right)a_R - a_D}\,\Delta P_I$*	$P_R + \dfrac{\lvert a_D \rvert\,\alpha}{(1+\alpha)a_R - a_D}\,P_R$
Q_R'	$Q_R - \dfrac{\frac{P_R}{P_I}}{\left(1+\frac{\Delta P_I}{P_I}\right)a_R - a_D}\,\Delta P_I$*	$Q_R - \dfrac{\alpha}{(1+\alpha)a_R - a_D}\,P_R$
P_I'	$P_I + \Delta P_I - \dfrac{a_I \frac{P_R}{P_I}}{\beta\left(\left(1+\frac{\Delta P_I}{P_I}\right)a_R - a_D\right)}\,\Delta P_I$*	$(1+\alpha)\,P_I - \dfrac{\alpha\, a_I}{\beta\left((1+\alpha)a_R - a_D\right)}\,P_R$
Q_I'	$Q_I - \dfrac{\frac{P_R}{P_I}}{\beta\left(\left(1+\frac{\Delta P_I}{P_I}\right)a_R - a_D\right)}\,\Delta P_I$*	$Q_I - \dfrac{\alpha}{\beta\left((1+\alpha)a_R - a_D\right)}\,P_R$

*Denotes approximation valid for small ΔP_I.

As Chapter 7 will argue, the import supply curve is relatively flat, so a_I is small compared with a_R and $|a_D|$. One reason for this is that there is practically an infinite supply of drugs outside the United States, and there are many people willing to try to smuggle drugs into the country. If demand in the U.S. increased, temporarily bidding up the import price so that smugglers began making excess profits, then more people would start smuggling until the import-export price difference were bid down to its equilibrium level. In other words, there are no appreciable diseconomies of scale due to constrained resources, so there is no reason for the import supply curve to have a steep slope.[20]

[20] Moore, 1986.

If a_I is indeed small, $\Delta P_I \approx \Delta\widehat{P}_I$, $\Delta P_I \approx \alpha\, P_I$, and $\Delta\widehat{P}_I \approx \widehat{\alpha}\, P_I$. Then, for small α, for the model used in the two previous studies

$$P_R' = P_R + (1 + rT)\, \Delta\widehat{P}_I \tag{3.22}$$

$$Q_R' = Q_R - (1 + rT)\, \frac{\Delta P_I}{|a_D|}, \tag{3.23}$$

for the additive model,

$$P_R' \approx P_R + c_1\, \kappa\, \Delta\widehat{P}_I \tag{3.24}$$

$$Q_R' \approx Q_R - c_1\, \kappa\, \frac{\Delta P_I}{|a_D|}, \tag{3.25}$$

and for the multiplicative model,

$$P_R' \approx P_R + c_2\, \frac{P_R}{P_I}\, \Delta\widehat{P}_I \tag{3.26}$$

$$Q_R' \approx Q_R - c_2\, \frac{P_R}{P_I}\, \frac{\Delta P_I}{|a_D|}. \tag{3.27}$$

For small changes in the import supply curve c_1 and c_2 are both approximately equal to $\frac{|a_D|}{a_R - a_D} < 1$. The elasticity of demand is probably lower than the elasticity of supply, so $|a_D| > a_R$, and thus as a crude approximation, $c_1 \approx c_2 \approx 1$.

3.8 The Three Models' Predictions About Prices

The previous section derived expressions for the new equilibrium retail price and quantity when the import supply curve shifts under the following assumptions: (1) the supply and demand curves are approximately linear over the range of interest; (2) the domestic distribution network retails a fixed percentage of the total quantity imported; and (3) the import supply curve shifts up by a constant amount for all quantities or by a constant percentage for all quantities. Then for modest changes in the import supply curve, the model used in the studies by Reuter and Kleiman (1986) and Reuter, Crawford, and Cave (1988) predicts that

$$P_R' = P_R + (1 + rT)\,\Delta\widehat{P}_I \tag{3.28}$$

$$Q_R' = Q_R - (1 + rT)\,\frac{\Delta P_I}{|a_D|}, \tag{3.29}$$

the additive model predicts that

$$P_R' = P_R + \kappa\,\Delta\widehat{P}_I \tag{3.30}$$

$$Q_R' = Q_R - \kappa\,\frac{\Delta P_I}{|a_D|}, \tag{3.31}$$

and for the multiplicative model,

$$P_R' = P_R + \frac{P_R}{P_I}\,\Delta\widehat{P}_I \tag{3.32}$$

$$Q_R' = Q_R - \frac{P_R}{P_I}\,\frac{\Delta P_I}{|a_D|}. \tag{3.33}$$

Since both (1 + rT) and κ are close to unity, the multiplicative model's predictions differ from those of the additive model and the model used in the previous studies by the ratio of the retail to import level price, $\frac{P_R}{P_I}$. Retail prices per pure unit for cocaine and heroin are considerably greater than their import prices. Hence, for a given increment in interdiction effort, the multiplicative model predicts that the resulting change in retail price and quantity consumed will be much greater than the changes predicted by the additive model. The numbers for high-level enforcement are similar although less extreme because $\frac{P_W}{P_I} < \frac{P_R}{P_I}$.

To be more specific, the model used in the two previous studies and the additive model predict that when prices change at one level of the domestic distribution network, prices at lower levels will change by approximately the same amount. That is, price changes are passed along dollar for dollar. In contrast, the multiplicative model predicts that the prices at lower levels will change by the change at the higher level **times** the ratio of the price at the lower level to the price at the higher level. That is, price changes are passed along on a percentage basis. Doubling the price at one level leads to a doubling of prices at all subsequent levels. The next section will examine which of these predictions more closely describes historical price trends.

3.9 Empirical Evidence

The predictions of the additive and multiplicative models are so different one might think that just looking at the import and retail prices of a particular drug in different years would show whether the additive model or the multiplicative model is more accurate. Indeed, that might be the case if one could simply look up "the" import and "the" retail price. Unfortunately, because of the inherent difficulties associated with collecting data on illegal activities, it is not that simple.

In the first place there are essentially no data on import prices.[21] Instead wholesale and retail data will be used. This makes the difference between the additive and multiplicative models' predictions less extreme, because the ratio of retail to wholesale prices is less than the ratio of retail to import prices. For cocaine purity adjusted wholesale prices are about four times higher than retail prices. For marijuana the ratio is about 1.5:1 for sinsemilla and 2:1 for commercial grade. For heroin the ratio is higher, but as will be explained in Subsection 3.9.3, the heroin data cannot be used to validate the model.

The price data are from the Drug Enforcement Administration's Office of Intelligence. (See Section 2.6.) These are essentially the prices reported by the National Narcotics Intelligence Consumers Committee (NNICC). Usually a range of prices is given, and the range can be quite broad. The mid-points of these ranges were used as a point estimate of the price.

The estimates were adjusted for inflation and purity. Prices were converted to 1989 dollars using the consumer price index.[22] (See Table 3.3.)

[21] Reuter, Crawford, and Cave (1988, p.80) note that import price data are not available. Although they use the term import price in their report, they actually use DEA data for large, domestic cocaine and marijuana transactions. The data below are from the same source.

[22] Economic Report to the President, Transmitted to the Congress February 1990, Table C-58, p.359.

Table 3.3:
Inflation, As Measure by the Consumer Price Index, 1982-1989

Year	1982 - 1984 = 100	1989 = 1.0
1982	96.5	0.7782
1983	99.6	0.8032
1984	103.9	0.8379
1985	107.6	0.8677
1986	109.6	0.8839
1987	113.6	0.9161
1988	118.3	0.9540
1989	124.0	1.0000

The purity adjustments were different for different drugs. The next subsection describes the evidence from cocaine price trends. The following subsection gives the corresponding evidence from marijuana prices. Subsection 3.9.3 describes why the price data for other drugs, including heroin, could not be used to check the models' predictions.

3.9.1 Cocaine Price Trends

The purity adjustment for cocaine was complicated but essential because there are substantial differences between wholesale and retail purities, and retail purities rose dramatically between 1982 and 1989. I asked Maurice Rinfret, in the DEA's Office of Intelligence, about purities of cocaine at the wholesale and retail levels.[23] He thought that purity has been essentially constant at the wholesale level throughout the 1980's; for a single number he would pick one between 87-88%. I used 87.5%. This number is fairly stable because lower purities mean smuggling larger quantities, but higher purities are difficult to obtain.

For retail purities he referred me to the GAO's report, *Controlling Drug Abuse: A Status Report*, which contains the official DEA estimates for 1981-1986.[24] He suggested augmenting those numbers with 55-65% for 1987 and 70% for 1988. Table 3.4 shows these purities.

[23] Telephone conversation, March 13, 1989.

[24] The numbers appear in a graph. It is difficult to be precise reading numbers from a graph, but it appears that they are all multiples of 2.5%, and even if some of the numbers are off by plus or minus 1%, it would not affect the argument.

Table 3.4:
DEA/GAO Retail Purity Estimates for Cocaine, 1981-1988

Year	Purity
1981	27.5%
1982	32.5
1983	35.0
1984	35.0
1985	55.0
1986	57.5
1987	55-65
1988	70.0

To be blunt, these purity data look suspicious. They show a dramatic jump in purity from 35% to 55% between 1984 and 1985. It seems more likely that there was a steady increase, suggesting that the DEA underestimated the rise in purity before 1984.[25]

One can easily imagine a scenario in which the experts discounted early reports that purities were rising rapidly. From the perspective of 1990 this seems foolish, but things looked different in 1984. Historically retail cocaine prices were low, so even the small increases reported before 1984 may have seemed large by the standards of the day. Perhaps in 1985 the experts realized that the reports from the field were accurate and representative, and they quickly adjusted their estimates, giving the sharp increase shown in Table 3.4.

At any rate I decided to look elsewhere for retail purity data. The only other data I could obtain was from the DEA's STRIDE system (described in Subsection 2.3.3). Jack Homer, who has been doing systems dynamics research on cocaine markets, used STRIDE data, and had done the considerable work of converting the data from the DEA's file system to an Excel Spreadsheet. Fortunately he was kind enough to give me his data.

In his work he used the average[26] retail purity observed in the between 237 and 874 records per year of seizures involving 0-6 grams per seizure, excluding records for which the cost was listed as 0. Figure 3.5 shows graphically the differences between the two sets of purity data.

Homer's purity data were for 1977 to 1987. I somewhat arbitrarily augmented these with my own extrapolations for 1988

[25] Reuter (1984) criticizes the DEA's monitoring of the retail trade.

[26] The average was a weighted average using the literal weights of the seizures.

and 1989. It is generally believed that retail purities have continued to increase, but there is an obvious upper bound to purity, so I assumed that the rapid rate of increase observed between 1982 and 1987 has slowed. Table 3.5 shows the retail purities used to test the models.

Figure 3.5:
Two Estimates of Retail Cocaine Purity Over Time

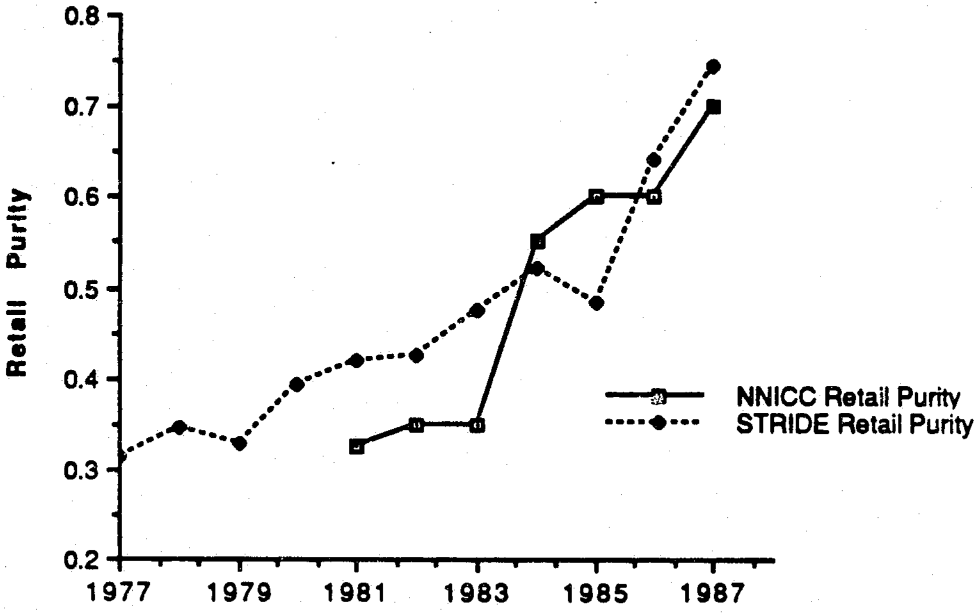


Table 3.5:
Estimates of Retail Cocaine Purity Used to Test the Models

Year	Purity
1977	31.5%
1978	34.6
1979	32.9
1980	39.4
1981	41.9
1982	42.6
1983	47.6
1984	52.3
1985	48.4
1986	64.2
1987	74.4
1988	75.4
1989	76.4

The purity and inflation adjusted cocaine prices are displayed in Table 3.6. These prices are the midpoints of the ranges of prices displayed in Table 2.12 divided by the inflation adjustment given in Table 3.3 and the purity adjustment (0.875 for wholesale prices and as given by Table 3.5 for retail).

Table 3.6:
Purity and Inflation Adjusted Cocaine Prices, 1982 - 1989
(1989 Dollars Per Pure Gram)

	National				
Wholesale (1 kg)	Range	Miami	New York	Chicago	L.A.
1982	$88.11	$78.57	$84.44	$91.78	$91.78
1983	71.14	39.13	56.91	71.14	71.14
1984	61.38	35.46	51.15	61.38	51.15
1985	52.68	42.80	48.73	55.97	49.39
1986	43.32	25.86	29.74	48.49	38.79
1987	32.43	16.84	28.07	37.42	17.46
1988	26.95	19.77	23.36	24.56	16.17
1989	23.43	17.43	21.71	20.00	15.43

	National				
Retail (1 gm)	Range	Miami	New York	Chicago	L.A.
1982	$339.34	$301.64	$301.64	$339.34	$377.05
1983	294.24	183.09	228.86	261.55	261.55
1984	251.01	159.74	199.67	228.19	228.19
1985	238.10	142.86	208.34	238.10	238.10
1986	176.23	96.93	149.79	176.23	176.23
1987	146.71	80.69	132.04	146.71	146.71
1988	118.16	97.31	97.31	121.64	104.26
1989	114.53	85.08	85.08	111.26	117.80

Figures 3.6 - 3.10 plot the retail prices against the wholesale prices. They show a surprisingly regular pattern. Retail price changes were proportional to wholesale price changes. If one labels the points with their dates, one also sees that cocaine prices declined steadily and substantially during the period. In Figure 3.6, giving the national range data, the earliest data point is in the upper right. Successive years' data move down the line to the lower left, ending with the most recent data point. Over the period wholesale and retail prices fell to about one-quarter to one-third of their original value.

Figure 3.6:
Retail vs. Wholesale Cocaine Prices National Range, 1982-1989
(1989 Dollars per Pure Gram)

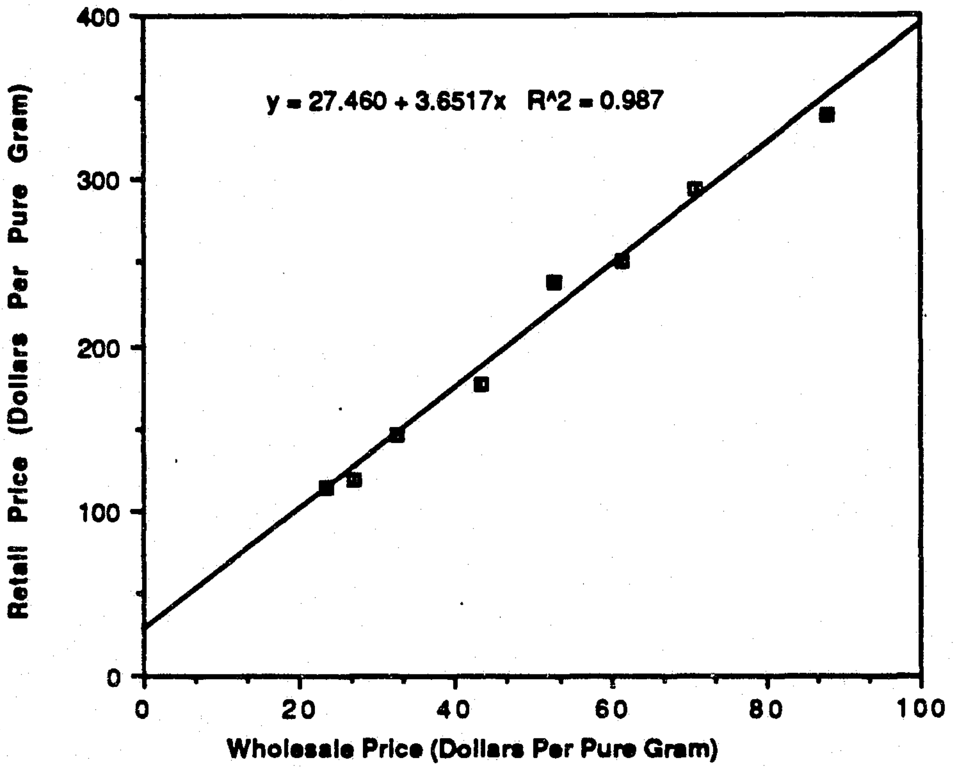


Figure 3.7
Retail vs Wholesale Cocaine Prices
Miami, 1982-1989

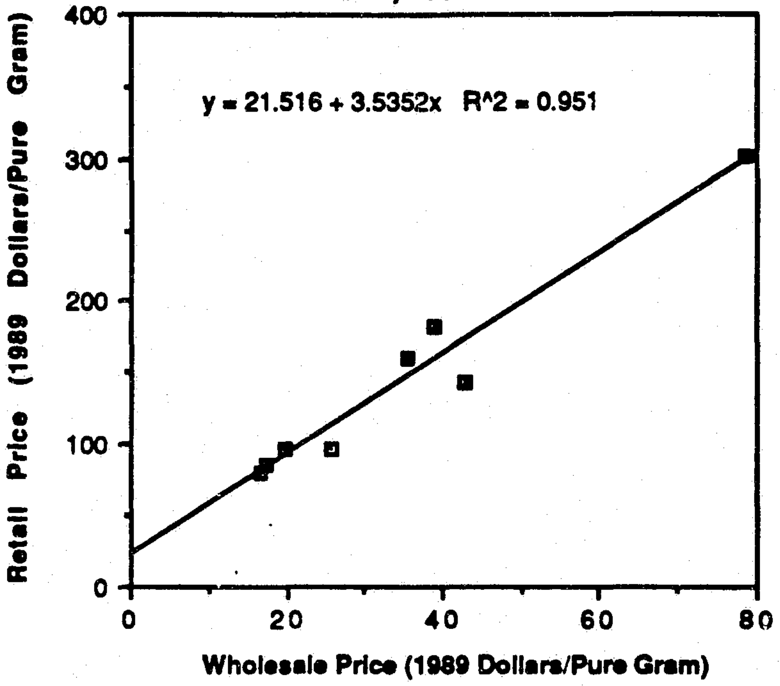


Figure 3.8
Retail vs. Wholesale Cocaine Prices
New York, 1982-1989

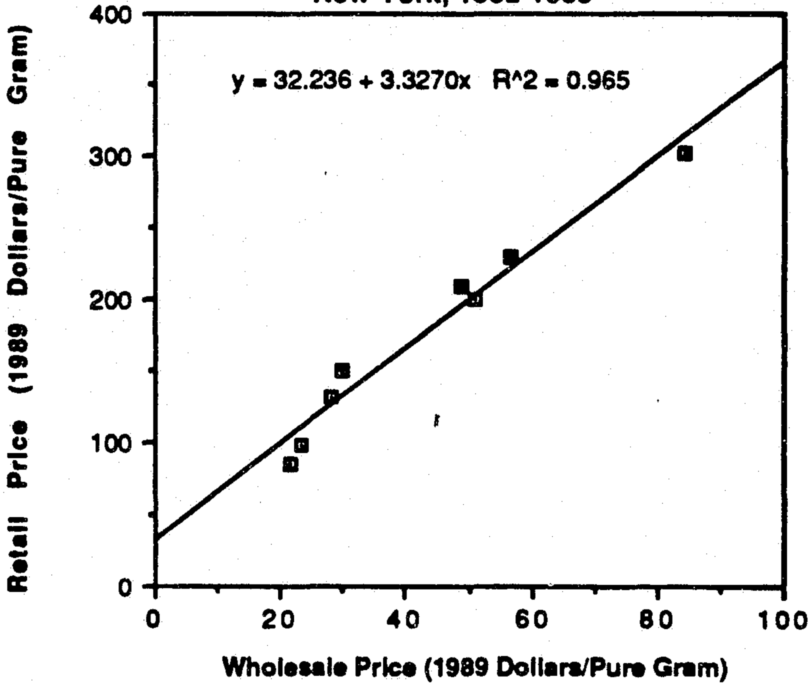


Figure 3.9
Retail vs. Wholesale Cocaine Prices
Chicago, 1982-1989

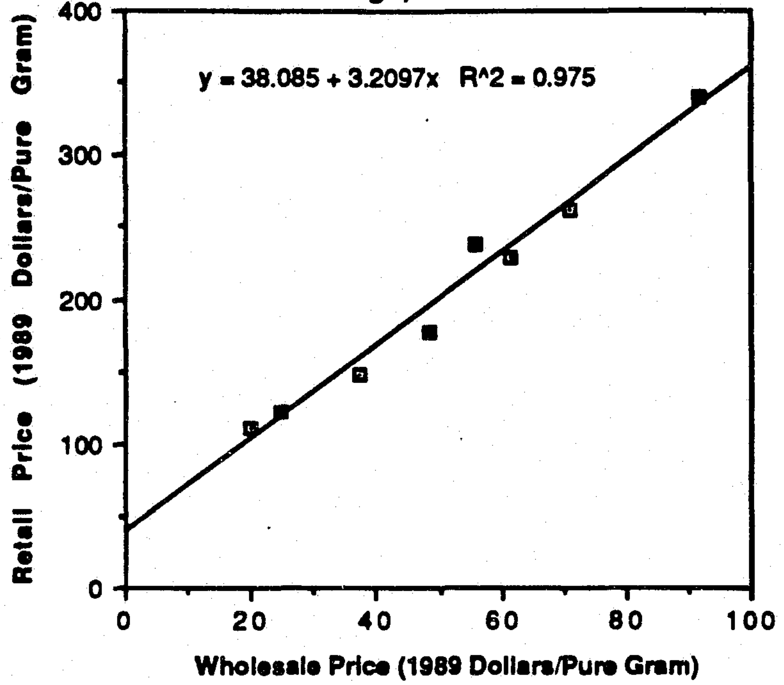
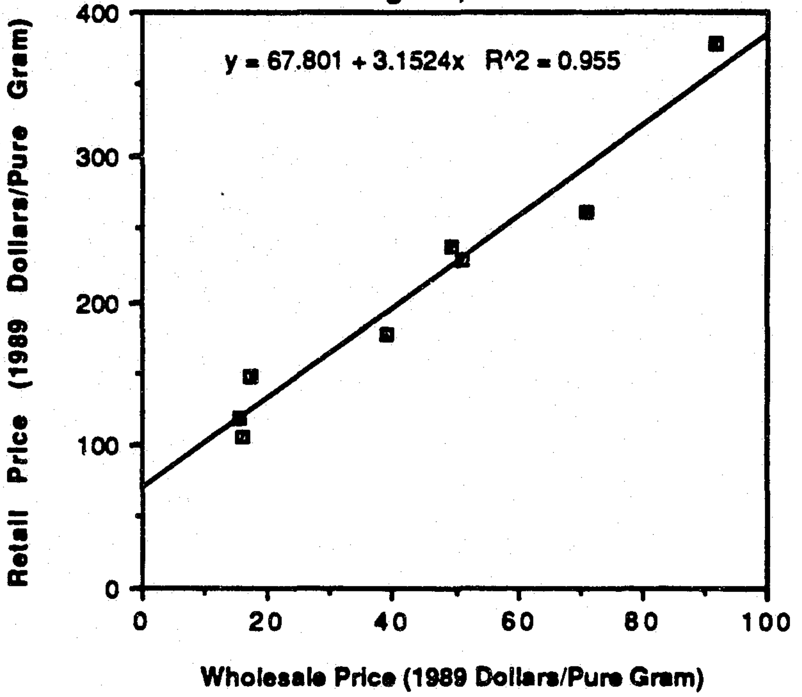


Figure 3.10
Retail vs. Wholesale Cocaine Prices
Los Angeles, 1982-1989



It appears that retail price changes are proportional to wholesale price changes, but the key question is, what is the proportionality constant? The additive model and the model used in previous studies predict that it should be close to one; the multiplicative model predicts that it will be slightly less than the ratio of the retail price to the wholesale price.

In Figures 3.6 - 3.10 the ratio of the retail to wholesale price changes, as measured by the slope of the least-squares line drawn through the data, is considerably greater than one, but it is also smaller than the ratio of the retail to wholesale price. (See Table 3.7.) Clearly neither model is perfect. One interpretation is that the true situation is intermediate between the additive and multiplicative models. The discussion in Section 3.5 suggested that the additive and value-preserving models were two extremes. The multiplicative model is an intermediate model, but it is closer to the value-preserving model. So perhaps the most accurate model would be intermediate between the additive and multiplicative models. That is, it would predict that the retail prices are more sensitive to changes in the import price than the additive model predicts, but less sensitive than the multiplicative model predicts.

Another interpretation is that the whole analytical framework developed here is limited. That is not to say that it is useless, but rather that the real world is complex and cannot be reduced to a simple model of the form described here.

Table 3.7:
Retail/Wholesale Price Ratio

	National Range	Miami	New York	Chicago	Los Angeles
Slope of line	3.65	3.54	3.33	3.21	3.15
Price Ratio in:					
1982	3.85	3.84	3.57	3.70	4.11
1983	4.14	4.68	4.47	3.68	3.68
1984	4.09	4.50	3.90	3.72	4.46
1985	4.52	3.34	4.28	4.25	4.82
1986	4.07	3.75	5.04	3.63	4.54
1987	4.52	4.79	4.70	3.92	8.40
1988	4.38	4.92	4.17	4.95	6.45
1989	4.89	4.88	3.92	5.56	7.63

It is clear from the table that the ratio of retail to wholesale prices has been increasing. Another way to see the same thing is to note that in all five figures there is a significant positive intercept to the least-squares fit line. One explanation for this is the following.

When the price of a drug is high, more of the costs and probabilities discussed above are likely to depend on the value of the transactions; when prices are low, more of the costs depend on the quantity. One way to see this is to think of licit goods. The cost of distributing jewelry depends more on its value than its weight, at least as compared with the cost of distributing cement. In general, the greater the price/unit weight, the less distribution costs will depend on weight (quantity). Hence one would expect that the higher the price of the drug, the more the price changes would follow the pattern suggested by the multiplicative model. The lower the prices, the more likely they are to follow the pattern suggest by the additive model.

Another view is purely descriptive. One could interpret the graphs as suggesting that there are fixed and variable costs. In this case the "fixed" costs depend on the quantity, but they are independent of price. The "variable" costs are costs that increase with price. At higher prices, the variable costs dominate and the price trends look like those predicted by the multiplicative model. At lower prices the "fixed" costs dominate, and the ratio of changes in the retail price to changes in the wholesale price starts to deviate from the ratio of the prices.

In as much as these views are accurate, one would expect that marijuana prices would follow the predictions of the additive model and heroin prices the predictions of the multiplicative model. The next section will show that, although the data are so poor it is difficult to conclude much, the marijuana data seem to more closely follow the pattern suggested by the additive model.

3.9.2 Marijuana Price Trends

The marijuana price data are less conclusive than the cocaine price data. Table 3.8 displays the inflation and purity adjusted midpoints of the price ranges given in Table 2.13 for commercial grade marijuana and sinsemilla.[27] Note, marijuana is not diluted, so the purity is the same at the retail and wholesale levels. The

[27] Commercial grade marijuana price data for 1982 and 1983 were separated into domestic, Mexican, Jamaican, and varieties. (See Table 2.13.) The prices in Table 3.8 are the average of the midpoints of the ranges for the first three. Colombian commercial grade marijuana price data are not used because the range of prices given for 1983 is much broader than any of those for 1982 or any of the other 1983 ranges.

adjustment is made only to facilitate comparisons between years and between commercial grade and sinsemilla.

Table 3.8:
Inflation and Purity Adjusted Marijuana Prices
(1989 Dollars per Gram of THC)

	Commercial Grade		Sinsemilla	
Year	Wholesale	Retail	Wholesale	Retail
1984	$34.80	$66.82	$72.34	$93.83
1985	30.91	82.40	55.83	89.35
1986	39.20	95.58	41.37	70.93
1987	62.60	105.72	52.84	89.38
1988	68.43	142.59	52.08	89.91
1989	68.35	114.01	66.89	91.42

The marijuana data are inferior to the cocaine data in three respects. First, there is less of it; the data only cover 1984-1989. Second, the wholesale level marijuana data is based on transactions of a pound of marijuana; retail data is based on ounce transactions. These are not that different, so there is not much difference between the wholesale and retail prices. Since the basis for distinguishing the models is the ratio of retail to wholesale prices, this limits the ability of the data to resolve the models. Finally, wholesale sinsemilla prices did not change much, so any relation between changes in retail prices and changes in wholesale prices could be masked by noise in the data.

Figures 3.11 and 3.12 plot the retail vs. wholesale prices. The ratio of retail price change to wholesale price change for commercial grade marijuana was 1.3:1 which is closer to the 1:1 predicted by the additive model than the ratio of retail to wholesale price (1.6:1 - 2.6:1) predicted by the multiplicative model.

The ratio of retail to wholesale price changes was actually less than one for sinsemilla, but the price changes were smaller and did not follow a consistent trend in either direction. So although one might argue that the sinsemilla data support the additive model, it is probably safer not to place much emphasis on them.

Figure 3.11
Retail vs. Wholesale Prices of Commercial Grade Marijuana, 1984-1989

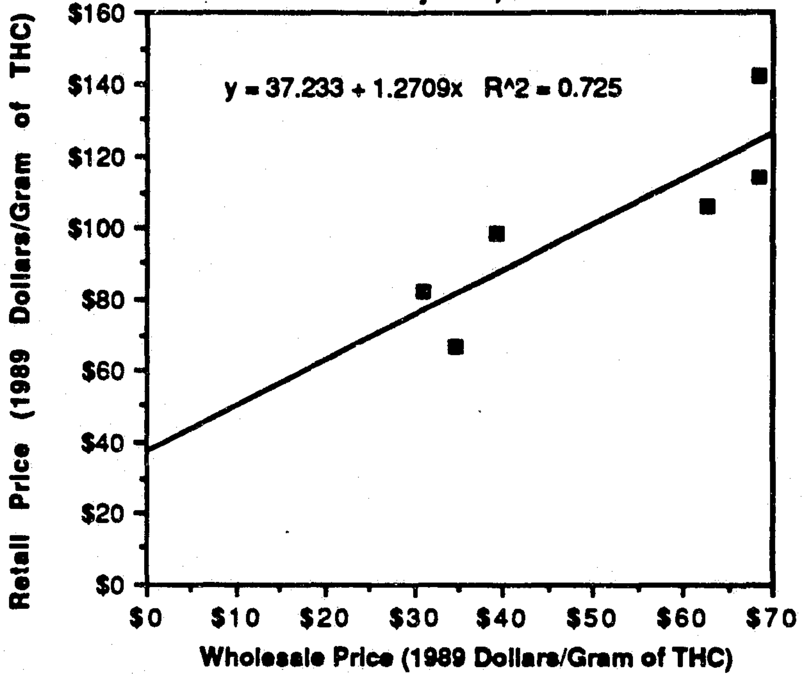
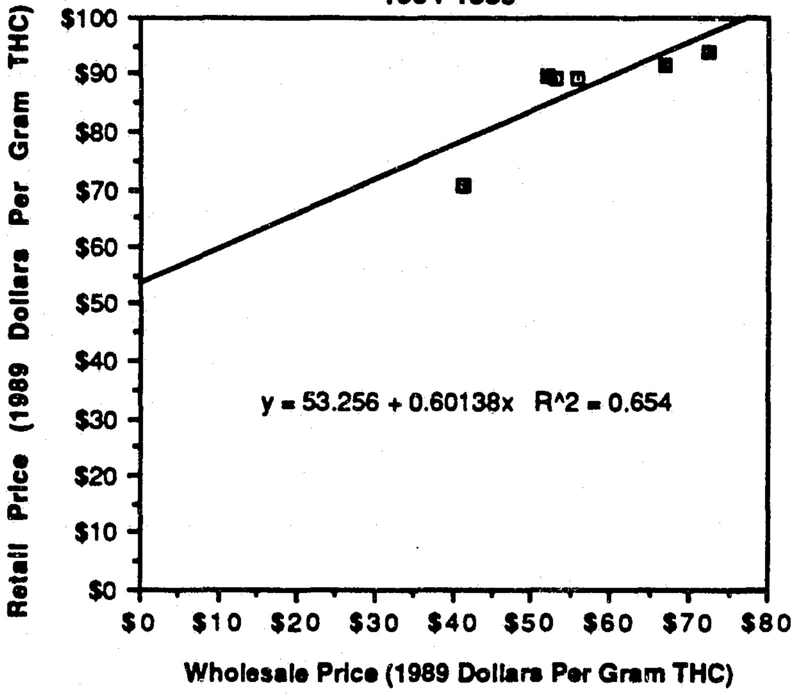


Figure 3.12
Retail vs. Wholesale Prices of Sinsemilla 1984-1989



3.9.3 Failings of Price Data for Other Drugs

The theory above is most applicable for drugs that are imported and distributed through long distribution chains, so heroin is the other drug with which one would like to test the models. The heroin price data provided by the DEA cannot be used, however, for a variety of reasons.

(1) The transaction quantities for which retail data are reported are different for the 1982-1984 and the 1985-1989 data.

(2) Wholesale (kg) data are divided between Southeast Asian, Southwest Asian, and Mexican heroin. Wholesale/retail (ounce) data, which are available only for 1985-1989, are not divided at all. Retail data for 1982-1984 are divided by geographic location (3 cities and the national range) and purchase type ("street quarter" or "dime bag"). The 1985 data are divided only by purchase type. The 1985-1989 data are divided by type of heroin (unspecified or "Mexican tar"). It is not clear how to match the data to compare trends.

(3) The only purity data given is for the wholesale/retail data between 1985-1989, and that data has considerable variability (40-80% in 1985; 10-70% in 1986; 30-80% in 1987; 20-80% in 1988; and 15-85% in 1989).

(4) The prices given cover a wide range. (The most extreme examples being the wholesale price for Southwest Asian heroin in 1989 is between $50,000 and $220,000 per kilogram and the retail price for a street gram in 1989 was between $60 and $300.)

The data for most other drugs offer even less hope for testing the models. For many there is no distinction between retail and wholesale prices. For others, such as Qualludes and PCP, the prices have been almost constant. For still others, such as Diazepam and LSD, prices have changed, but there was no trend. The changes can best be described as an increase in the range of prices observed at both the wholesale and retail levels. The methamphetamine data would seem to offer the best hope, but there has been no steady trend in methamphetamine prices. Wholesale and retail prices have both moved up and down, but in any given year they are as likely to have moved in the opposite direction as in the same direction.

3.10 Summary

Recently there has been considerable debate about the proper division of resources between controlling supply and reducing demand. Some people argue that border interdiction and high-level enforcement are futile. They believe that the markets operate too smoothly for them to reduce availability and that increasing the price at the upper-levels of the distribution chain will not appreciably affect consumption because the import and high-level wholesale prices are only a small fraction of the final retail price.

Implicit in this argument is the assumption that the domestic distribution network passes along price increases on a dollar for dollar basis, as would be the case for most licit goods. This chapter offers an alternate model of the price linkage; it suggests that price increases are passed along on a percentage basis. A 10% increase in the import price would lead to a 10% rise in the price at all subsequent levels. These two views are labeled the additive and multiplicative models, respectively.

A decision analytic argument is used to suggest conditions under which the additive model and another model called the value-preserving model might hold. Conditions in the real world probably fall somewhere between these two sets of conditions, suggesting that a compromise between the additive and value-preserving models may be the most successful in explaining historical data. The multiplicative model is one such compromise, although it is more like the value-preserving model than the additive model.

For a variety of reasons it is difficult to test the models with historical price data. The test performed with marijuana data lent some support to the additive model. Comparisons with cocaine price data seem to support the multiplicative model.

One possibility is that the higher the overall price level, the more the price linkages follow the predictions of the multiplicative model, and the lower the price level, the more they follow the predictions of the additive model. Hence in the early 1980's, the price linkage for cocaine was very much like that predicted by the multiplicative model. On the other hand, for marijuana and to a lesser extent for cocaine in the late 1980's, the linkages may be more like those predicted by the additive model.

Regardless of which model is better, the price data for cocaine are themselves interesting. When the data are adjusted for changes in purity and inflation, they reveal that retail and wholesale cocaine prices moved almost in lock step between 1982 and 1989. At the national level the very nearly followed the equation

$$\text{Retail Price}_t = \$27.50 + 3.5 \text{ X Wholesale Price}_t \qquad (3.34)$$

where prices are measured in 1989 dollars per pure gram.

Assuming this relation is causal, that is, that changes in wholesale prices cause changes in retail prices, then the argument against interdiction and high-level enforcement given above may not be valid. It may be that increasing prices near the top of the domestic distribution network would significantly increase retail prices, thereby reducing consumption.

One might ask why, if this is true, did retail prices fall even as the Reagan Administration stepped up interdiction efforts? The answer may be simply that even though expenditures on interdiction increased, the import price actually fell, perhaps because international supply increased or because smugglers' skills improved even more dramatically than those of the interdiction agencies.

So this chapter by no means argues that interdiction and high-level enforcement are a panacea. Even if it is true that forcing up prices high in the network increases retail prices, one must be able to increase those higher-level prices for interdiction and high-level enforcement to work, and that does not seem to be possible at this time.

CRACKDOWNS: A MODEL OF ONE FORM OF LOCAL DRUG ENFORCEMENT

Jonathan P. Caulkins

Chapter 4: Crackdowns: A Model of One Form of Local Drug Enforcement

4.0 Introduction

Crackdowns are being promoted as a new approach to drug enforcement.[1] What exactly are "crackdowns"? Kleiman[2] discusses them at length, but for the purposes of this chapter a working definition is: "an intensive local enforcement effort directed at a particular geographic target." Frequently the target is a so-called "open-air" drug market, where there is a high concentration of dealing and the market participants fairly flaunt their presence.

The definition seems so general one might ask what distinguishes a crackdown from other enforcement operations. Two things do.

First, crackdowns involve concentration of resources. In contrast, day-to-day enforcement operations often spread resources more or less uniformly over the "problem". This may be done to keep any one part of the problem from getting completely out of control and for equity considerations.

Second, crackdown targets are geographic. Other kinds of drug operations may target individuals, organizations, a particular drug, a class of users, or an ethnic group.

Although local-level enforcement[3] in general and crackdowns in particular have been receiving considerable attention lately, there is no consensus about how well they work. This chapter attempts to contribute to the debate by developing a mathematical model of crackdowns.

The next section reviews some current arguments for and against local enforcement and crackdowns. Section 4.2 briefly describes the situation in Hartford, Connecticut, a city that is undertaking a crackdown program. The author spent the summer of

[1] For example by Hayeslip, 1989. Moore and Kleiman (1990) describe seven strategies police can use against the drug problem; one of them is "neighborhood crackdowns."

[2] Kleiman, 1988a.

[3] Local-level enforcement generally refers to operations conducted by police and sometimes sheriffs departments. They usually focus on individuals and organizations that deal within a small region, frequently a city. In contrast, high-level enforcement operations are usually conducted by federal agencies and target importers, wholesale dealers, and organizations that cross local agencies' jurisdictional boundaries.

1989 riding with two narcotics detectives in Hartford, so the model's assumptions and structure reflect Hartford's situation and needs. As is discussed, Hartford shares many characteristics of large American cities, so it is hoped that the model's implications apply more generally.

Section 4.3 describes two mental models that inspired the more formal mathematical model introduced in Section 4.4. Section 4.5 solves the model for three special cases that presage some results that hold more generally.

Section 4.6 describes results that hold for general parameter values; it contains the most important mathematical results. Section 4.7 presents explicit solutions for some sets of parameter values.

Sections 4.8 - 4.13 describe various applications of the model and related issues. Section 4.14 summarizes the model's conclusions and its implications for Hartford. The last section considers which of the model's results can be extrapolated to the national market.

4.1 Current Thinking About Local Enforcement

This section briefly discusses some of the advantages and disadvantages of local-level enforcement and crackdowns.

4.1.1 The Promise of Local-Level Enforcement

One cynical explanation for the interest in local-level enforcement is that the Reagan Administration emphasized interdiction and high-level enforcement,[4] but both seem to have "failed". So those who think the status quo is unsatisfactory but are unwilling to join the legalization camp need to find another remedy.

The enthusiasm for local-level enforcement is not all Pollyannish though; there are sound theoretical reasons for supporting local enforcement. One of the most compelling is that it can increase search time costs.[5] Like a price increase, increasing search costs raises the overall cost of acquiring drugs and presumably that reduces consumption. Furthermore, the cost increase, and hence the strength of the disincentive, may be greatest for novices, and reducing experimentation and recruitment are particularly valuable.

[4] The emphasis on "supply control" rather than "demand reduction" began as early as 1977, but it became much more pronounced during the Reagan Administration (Marshall, 1988a).

[5] See Moore (1973) and Kleiman (1988a).

Increasing search time costs has a different effect on overall spending for drugs, however, than does an increase in their dollar price. When the price of a good, even an illicit good, goes up, people almost always consume less of it.[6]

For some goods relatively small changes in price lead to large changes in the quantity consumed. A good is called price elastic if the percentage change in quantity exceeds the percentage change in price (for small price changes). When the price of one of these goods increases, spending on that good declines.

Other goods are what is called price inelastic. When the price of these goods increases, people consume less of them, but the percentage decrease in consumption is less than the percentage increase in price, so spending on that good increases.

Measuring elasticities is difficult even for licit goods, so the best one can hope to do for illicit goods is to make educated guesses. The experts who have done this generally agree that demand is inelastic, but the quantity does respond somewhat to price so it is not perfectly inelastic.[7]

Hence one would expect that increasing the price of drugs would increase spending on those drugs. This has a number of undesirable consequences. It increases drug dealers' revenues; it impoverishes users, many of whom can ill-afford to divert spending from other items; and, in as much as a portion of drug purchases are financed by property crime,[8] it could well lead to an increase in property crime. According to Kleiman,[9] "The one empirical study addressing this question suggests that increasing heroin prices tend to generate increases in property crime, but the question is far from settled."

In contrast, local-level enforcement could decrease spending on drugs and hence decrease dealers' revenues, allow users to spend more on other goods, and reduce property crime. The hope of reducing crime is one of the principal arguments given in favor of local enforcement.

[6] The exceptions, called Giffen goods, are rare.

[7] See, for example, Reuter, Crawford and Cave (1988, pp.20-23) and Reuter and Kleiman (1986, pp.298-300).

[8] A causal relationship has not been proved in the scholarly sense of the word, but the existence of documents such as "Reducing Crime by Reducing Drug Abuse: A Manual for Police Chiefs and Sheriffs" (International Association of Chiefs of Police, 1989) demonstrates that practitioners have accepted it as a working principle.

[9] Kleiman (1988a, p.20), referring to Brown and Silverman (1974).

Kleiman[10] gives evidence that property crime may in fact have been reduced by crackdowns in Lynn, Massachusetts and in New York City. As Barnett argues, however, the evidence is not conclusive, although he agrees that it justifies additional experiments with local-level enforcement.[11]

4.1.2 Prison Capacity: A Limitation on Local Enforcement

The main problem with local-level enforcement can be stated simply. Retail markets are huge. There are literally hundreds of thousands of dealers serving millions of customers,[12] and it is commonly believed that there are many people willing to take the place of any dealers the police remove. There simply is not room for them all in existing prisons,[13] and many people would object to incarcerating that many people even if there were room.

4.1.3 Crackdowns: A Way Around the Prison Capacity Constraint

One way to achieve some of the benefits of attacking local-level markets without swamping the criminal justice system is to focus on a (geographically) small target. This strategy, known as a "crackdown", has been tried in a about a dozen cities with varying degrees of success,[14] and is now being promoted for widespread use.

Crackdowns allow police to target the dealing which creates the worst externalities. Not all retail transactions are equally "bad" in the sense that the societal cost per gram or per dollar can vary considerably.

Open-air drug dealing imposes particularly large costs.[15] It advertises and glamorizes drug use and drug dealing, makes it easier for both novices and experienced users who have just moved to the area to score, disrupts the community, and engenders more dealer-dealer violence than does, for example, the "quiet" dealing that

[10] Kleiman, 1988a.

[11] Barnett, 1988.

[12] Reuter and Kleiman (1986, p.294) give 725,000 as a rough estimate of the combined number of retail heroin, cocaine, and marijuana dealers.

[13] Prisons today are clearly crowded (Bureau of Justice Statistics, 1989), and local-level narcotics enforcement has been blamed for aggravating this situation (Pitt, 1989).

[14] Chaikan (1988) includes a discussion of some of the successes and failings of crackdowns.

[15] Evidence for this is that "police in many jurisdictions have been besieged with complaints from residents of neighborhoods where drug dealing and 'dope houses' operate," Hayeslip (1989, p.2).

occurs in the work place. Crackdowns typically focus on so-called open-air drug markets for these reasons and because it is easier for police to make arrests there than to break up "quiet" dealing.

If one pays attention to the popular press, one might think that the vast majority of retail sales occur on the street, but this is not the case. Past estimates of street sales' fraction of total sales ranged from 25% to 90%,[16] but some recent evidence suggests it might be even less. Ninety-seven arrestees who self-reported crack use and were willing to give information reported that only 9.3% of their purchases were made from the "street" and another 9.3% from "touters".[17] "Dope houses" or "crack houses" accounted for two-thirds of their purchases. This is particularly significant because one might expect a sample of this type to purchase a larger than average fraction of their drugs on the street.

Improving the quality of life in the neighborhood is another reason for trying to shut down open-air drug markets. Most people do not want to live near open-air drug markets. They are noisy,[18] spawn violence that can affect innocent bystanders,[19] and may make it more likely that children in the neighborhood will become involved in dealing or using. District Attorney Burke thinks these factors are so important that he considers the Lawrence, Massachusetts crackdown to have been a success even though it apparently did not reduce drug-related street crime.[20]

Peter Reuter offers another, slightly different, argument for how eliminating open-air selling might reduce consumption.[21] He argues that street markets are the "7-11's" of the drug distribution system. He hypothesizes that many of the customers have alternate, regular sources, but that periodically they wish to supplement those sources.

For example, they may have their own supply at home, but want to buy immediately enough for a single use. Or perhaps they have exhausted their supply, and although they have arranged to meet their regular supplier in a few days, they want some now and cannot move up the meeting with their regular supplier.

[16] Garreau (1989) gives this range of estimates.

[17] Mieczkowski, 1989.

[18] Kleiman (1988a) reports that people living around Washington Park in New York City complained as much about the noise and general unpleasantness of dealing in their neighborhood as they did about violence or property crime.

[19] Daley and Freitag, 1990.

[20] Burke, 1988

[21] Reuter et al., 1990.

Whatever the reason, the customers want the equivalent of fast food service. If they could not obtain the drugs so conveniently, they might consume less.

For an analogy one might think of coffee consumption. People can buy coffee in the grocery store, make it at home, and bring it to work in a thermos. Instead, even though it is much more expensive, people frequently buy coffee a cup at a time in doughnut shops, company cafeterias, and similar institutions. Intuitively it seems plausible that if people could not buy a cup of coffee so conveniently, and instead had to bring it with them from home, they would drink less coffee. It may also be so with illicit drugs.

4.1.4 Problems With Crackdowns

The chief problem with narrowly focused crackdowns is the possibility of displacement. If the police close down one market, the customers and dealers may simply move to other, pre-existing markets without reducing their consumption, violence, or property crime. (They could conceivably move en masse to a new location that was not previously a center for dealing, but this is less likely because there is generally no way to coordinate their actions.)

One counterargument is that even if displacement occurs, concentrating dealing in a few neighborhoods might be desirable. Two of the most important benefits of closing down a drug market, improving the quality of life in the neighborhood and increasing search time by reducing the number of open-air markets, are still achieved even if there is full displacement.

Kleiman and Smith[22] offer an analogy with litter in public parks. Suppose there are 10 polluted parks, but the city only has the resources to pick up 10% of the litter. Removing 10% of the litter in each park is fair, but it still leaves 10 dirty parks. In contrast, cleaning up all the litter in one park accomplishes something tangible; it creates one clean park.

Issues of equity are more serious for drug markets than they are for clean parks. City residents can all visit the clean park. The residents of the neglected drug markets, however, do not derive much benefit from having a street somewhere else in the city freed of dealing.

Furthermore, in as much as there is displacement, the city park analogy is incomplete. The story would be a closer parallel if it finished with the city workers dumping the trash they gathered from the favored park in the less fortunate ones. That would almost

[22] Kleiman and Smith, 1989, pp.23-24.

certainly cause an outcry, and having drug dealing pushed into one's neighborhood is more cause for concern than an increase in litter.

On the other hand, simply pushing dealers and customers from market to market may offer some advantages that the litter analogy hides. For one, it disrupts connections. Customers and dealers in a market may establish a relationship which reduces uncertainty and search time. Then if their market is closed down, even if they both move to new markets, unless they happen to move to the same new market, that connection will be broken.

Closing down markets and displacing dealers may also make turf wars more frequent. Increasing dealer-dealer violence would increase the cost of distributing drugs and hence their retail prices, but it is not clear that the resulting reduction in consumption would be worth the additional violence.

Displacement is not the only problem with crackdowns; corruption is also a concern.[23] There are two kinds of corruption: practices that lead to the arrest and conviction of innocent people and practices, such as accepting bribes, that reduce the chance offenders will be prosecuted. Police Chief Bouza points out the first can happen if crackdowns pressure police to produce a large number of arrests. This can lead to "flaking, dropsy, perjury, entrapment, and framing."[24]

Bribery occurs with all types of drug enforcement, but seems more likely to happen with local enforcement for at least two reasons. First, the police officers and dealers usually know each other; they may have almost daily contact. In contrast, DEA and FBI agents rarely see the targets of their investigations (except perhaps during undercover operations). Second, DEA and FBI agents are less confined to a specific investigative target, so a dealer would have to bribe many agents to gain protection. In contrast, a relatively small number of police officers may be responsible for the area in which a particular low-level dealer operates, so the dealer would not have to bribe as many people.

Another problem with crackdowns is that they can be demoralizing for the officers involved. Making cases against users and low-level dealers is not professionally rewarding. To state it politely, minimal investigative skills are required. To put it bluntly, sticking one's hands in other people's pants all day long is demeaning

[23] For example, Bouza (1988) states that the New York City Police Department would swap entire vice units with uniformed patrols without warning to (largely unsuccessfully) break up corrupt practices.

[24] Bouza, 1988, p.48.

to everyone involved. (For obvious reasons that is where many dealers hold their drugs.) And frequently those arrested are back on the street within days if not hours.

Furthermore, street-level enforcement is dangerous. High-level investigations involve many hours of surveillance, sitting on wire taps, and "gum-shoe" work. The relatively low-level investigations local police undertake also have some of this character. In contrast, during crackdowns against street dealers police spend more of their time physically pursuing suspects and making arrests.

During the summer of 1989, narcotics detectives in the Hartford Police Department participated in raids almost once a day.[25] Every raid involved battering down a door and charging in with guns drawn. There were often firearms in the apartment, and although the police were never shot at that summer during such a raid, the people inside frequently resisted. "Street-rips"[26] and "Buy Busts"[27] are no safer; in fact, the detectives in Hartford think doing "buys" is more dangerous than going on raids.

Done properly, local enforcement can improve police-community relations,[28] but practice can differ from theory.[29] Many of the police are white, and many of the people who are hassled and arrested are minorities. The recent Stuart murder case in Boston shows the level of racial tension that can exist between white police and minority residents.[30]

Even if racism is not a factor, local-level enforcement can harm police-community relations. A distressing fraction of young males in the inner-city deal drugs.[31] Even if their families oppose their

[25] Personal observation.

[26] Attempts to break up dealing on the street without the benefit of prior intelligence.

[27] Buy busts are operations in which "undercover officers buy drugs on the street and then arrest the sellers" (Hayeslip, 1989, p.3).

[28] This is one of the goals of community policing.

[29] Kleiman (1988a) briefly describes the disastrous "Operation Cold Turkey" in Philadelphia which was quickly abandoned because of abuses and citizen hostility. According to Canellos (1990), "By the end of the crackdown, 1,444 people, most of them black, had been detained. In virtually none of the cases, it was later disclosed, did the police have cause to search them. ... None [of the 80 found to be carrying drugs] could be prosecuted because the crackdown was ruled illegal. ... Afterward, the city paid $500,000 to those illegally stopped."

[30] Described by Martz, Starr, and Barrett (1990) and Alter and Starr (1990).

[31] Reuter et al. (1990, p.vii) report that 16% of the black males residing in Washington D.C. who were born in 1967 where charged with selling drugs between the ages of 18 and 20.

dealing, the families may still resent the police for harassing and/or arresting a member of their family.

In short, crackdowns are not as glamorous as they sound. Street enforcement is ugly, violent, and can exacerbate racial tensions.

Higher-level implementation issues can also be problematic. For example, how does one select the crackdown target? This chapter tries to find some objective criteria for doing this, but if these or similar criteria are not followed, the process of choosing a target is prone to abuse.

Residents and local politicians may object vehemently for fear the crackdown will stigmatize their neighborhood. Or they may fight to have the crackdown if they believe it will actually help. Residents and politicians from adjoining neighborhoods may oppose the crackdown because they fear that it will displace dealing into their neighborhoods. At any rate, the choice of target may have more to do with the relative political power of various neighborhoods than with any objective criterion.

One way of partially circumventing this is to plan a crackdown campaign that cleans up all the markets in a city one at a time. This restores some equity, but it may still be advantageous to be near the top of the list, so the rank-order can be contentious. At least some displacement seems likely, so dealing might get worse before it got better in the markets that are not attacked first. Furthermore, if the police ran out of resources before they finished, they might replace ten relatively mild markets with two or three "combat zones." That may be good for the city as a whole, but it is almost certainly not good for the people who live in the newly created combat zones.

A final concern is that driving dealing indoors is not always unambiguously good. As mentioned above, open-air dealing probably generates more negative externalities per transaction than does "quiet" dealing, but not all indoor dealing is "quiet". The popular press is full of horrifying stories about so-called "open" crack houses where customers can use as well as buy drugs, and both crack houses[32] and shooting galleries[33] contribute to the spread of AIDS. If the alternative to street dealing is dealing in crack houses and

[32] There are three reasons for this. First, crack houses may sell regular cocaine as well as crack, some of which is injected. Second, apparently some users participate in sexual acts, including those with a high risk of transmitting the HIV virus, while they are using crack to heighten the pleasure (Power and Wells, 1989). Third, prostitution, for money or directly for drugs, also occurs in crack houses (Jacobs, 1989).

[33] See Chapter 6.

shooting galleries rather than so-called "quiet" dealing, it is harder to make the case that driving dealing indoors is a good thing.

In summary, there are arguments for and against local-level enforcement, including crackdowns. This chapter seeks to help inform this debate by introducing a mathematical model of crackdowns.

Like all formal models, this one is based on an intuitive understanding of the subject matter. The next two sections describe the basis for that intuition. The next section describes the city of Hartford, Connecticut where the author spent the summer of 1989 observing local enforcement operations and retail drug markets. The following section describes two mental models the embody some of those observations and the conventional wisdom about crackdowns.

4.2 Hartford, Connecticut

Despite the problems described above, cities continue to plan crackdowns.[34] Hartford, Connecticut is one such city. The model developed in this chapter was formulated with Hartford in mind because that is where the author learned about local enforcement and retail drug markets. So this section will briefly describes that city.

Hartford itself is quite small (population 138,000[35]), but it shares characteristics with larger cities for several reasons. The first of these is that the city limits encompass only a fraction of the people who live in the area. The Hartford Standard Metropolitan Statistical Area has a population of 726,114, which makes it the 55th largest in the country.[36] Furthermore, Hartford's population density (7,752 people per square mile) approaches that of cities like Los Angeles (6,996), Detroit (8,010), and Washington, D.C. (9,984).[37]

The core city of Hartford is very poor. Less than a quarter of its housing units are owner occupied; only half the residents graduated from high school; the mean family income ($16,580) is considerably less than the average personal income per person for

[34] Kleiman (1988a) notes that at the time of his writing, six cities had received funding for street-level enforcement from the Bureau of Justice Assistance.

[35] Unless otherwise noted, the information about Hartford given in this section is taken from Sullivan (1988, Chapter II).

[36] By the New England County Metropolitan Area definition Hartford is even larger, containing over a million people and ranking 34th largest in the country (U.S. Bureau of Census, 1987).

[37] U.S. Bureau of Census, 1987, Table 38.

the state as a whole;[38] over a quarter of the population live below the poverty line; and the crime rate is almost double the average for ten comparably sized cities.

And Hartford has a drug problem.[39] Strangely, very little crack is used, but cocaine and heroin abuse are widespread. As was discussed in Chapter 2, it is difficult to know even approximately how many users there are, but there are enough to enable the Police Department to identify 22 distinct drug markets in the city.[40]

Currently the Hartford Police Department has 30 people in its vice and narcotics division. Historically their efforts have been dispersed throughout the city, but there is a plan to concentrate efforts on one or two of the 22 markets. The long term plan is to clean up all 22 markets in succession. Then, assuming none of the original markets grow back and no new markets form, Hartford will have successfully driven drug dealing off its streets. Hence the Hartford Police Department needs to decide:

(1) How many and which of the 22 markets to attack first?
(2) How much pressure should be maintained on markets that have already been cleaned up when the main thrust goes on to other markets?
(3) When should the crackdown begin?

This chapter tries to give at least partial answers to these pragmatic questions as well as to the more theoretical questions raised in the next section.

4.3 Mental Models that Led to the Balloon Model

Much remains to be learned about crackdowns. This chapter tries to move toward a better understanding by formalizing two mental models that arose during discussions about crackdowns.

The first mental model is the "balloon metaphor."[41] To understand it, imagine a map of the city of interest with a sheet of rubber draped over it. The rubber is puffed up (hence the name "balloon" metaphor) over the points on the map corresponding to

[38] $19,600 in 1986 (U.S. Bureau of Census, 1987, Table 682).

[39] Described by Hohler, 1989.

[40] This may seem hard to believe until one finds out that Washington D.C. is thought to have 91 (Garreau, 1989).

[41] Richard C. Larson developed the balloon metaphor in the context of local drug enforcement during his work with the Hartford Police Department.

drug markets. The height of the balloon at any point is proportional to the density of dealing, and thus the volume under one bubble measures the size of that market.

The density of dealing might be measured in units such as dealers per block or dollar value of sales per acre. If there were two markets with the same amount of dealing but one was more geographically disperse, the balloon over that market would be lower and broader than the other; the second might look like a sharp peak if the dealing were highly concentrated. (Obviously a real balloon can not have the latter shape, but the language of the balloon metaphor is used because it is succinct and colorful.)

Another mental model (which is not confined to local enforcement) is the idea that enforcement can generate a "positive feedback effect."[42] As enforcement increases, some dealers who are particularly sensitive to enforcement pressure exit the market. That increases the amount of enforcement per participant among those who remain, which might encourage still more to leave. The departure of this second group, even if total enforcement pressure remains the same, further increases the ratio of enforcement to the size of the market.

If the market is small enough relative to the level of enforcement, this positive feedback effect might collapse the market. In effect, for any given level of enforcement there is a minimum viable market size.

These mental models suggest asking the following questions:

(1) Is there any advantage to focusing effort on one market?
(2) How hard does one have to push down to dent the market?
(3) If one pushes down hard enough, will the market pop (collapse)?
(4) If so, how much will it gradually deflate before it pops?
(5) If one pushes down hard enough to partially deflate the market, but not hard enough to pop it, will the market spring back?
(6) When a market is partially deflated or completely burst, is the dealing simply displaced to other markets or is it truly eliminated?
(7) If it is displaced, does it move only to adjacent markets or is it spread more or less uniformly over all the other markets?
(8) How much pressure is needed to keep a popped market from springing back?
(9) Is the effort required to pop a market proportional to its size? To the square of its size? To some other power of its size?
(10) What affects the proportionality constant?

[42] Discussed by Kleiman, 1988a, pp.25-26.

The model developed below to try to answer some of these questions is called the balloon model. The mental model described in this section is called the balloon metaphor to distinguish it from the more formal model developed next. It is important to remember that, although the two are similar and one was the inspiration for the other, they are different. The balloon model obviously has the advantage of being more precise, but it does not supersede the balloon metaphor for at least three reasons. First, the balloon metaphor is easier to explain and communicate. Second, the balloon metaphor's visual imagery is rich and could lead to further insights that equations hide. Third, the formal model developed next does not address directly inter-market interactions, including displacement.

Both the balloon model and the balloon metaphor are valuable, and the names for them were chosen to preserve this value by noting their similarities and their differences.

4.4 Model Formulation

Imagine a city with a large number of identical dealers spread over a large number of drug markets. Suppose that each day every dealer goes to the market that offers the dealer the best "opportunity". On any given day a dealer could also choose not to deal; to avoid constantly adding this qualification, not dealing will be thought of as going to one particular market, a "null" market.

Opportunity as used here is not synonymous with expected profits. It includes non-monetary factors such as the risks of enforcement and non-monetary costs (e.g. threats of violence) imposed by other participants in the market.

If dealers are identical, they share the profits and the burden of enforcement equally. Hence, all dealers in a market do equally well, and one can think of an expected return, or utility, from dealing in that market.

Furthermore, the utility would be the same in all markets that have any dealers. Suppose one market had a lower return. Then the next day fewer dealers would go to that market. If reducing the number of dealers increased utility, parity would be restored. If reducing the number of dealers decreased utility further, then still more dealers would leave, until eventually the market disappeared.

Realistically not all dealers can go to all markets. Some markets might be simply too distant; others may be inaccessible because the dealers and/or neighbors there are of a different ethnic background and would not welcome an outsider; and others may be on the "turf" of another gang. The model does not, however, require such extreme mobility. It assumes only that there is sufficient mobility to prevent dealers in one market from consistently earning higher returns than the dealers in other markets.

This is similar to the microeconomic assumption of free entry. Standard microeconomics assumes that in the long run, no industry can consistently produce higher profits than other industries because if it did, firms in other industries would move into the profitable industry. Clearly not all firms are equally capable of participating in all industries, but the assumption is that there are enough that are at least partially mobile that in the long run profits in different industries will be equal.

When all non-empty markets yield the same utility, city-wide dealing is in equilibrium. Dealers would have no incentive to change markets (or to start or stop dealing altogether).

Now suppose enforcement pressure changed, for instance by increasing enforcement in one of the many markets. Presumably the utility in that market would decrease. The markets would no longer all be in equilibrium, so dealers would consider changing markets.

If the change in enforcement were not too great, it might be that only a few dealers would leave that market. This would leave more customers to those who remain, compensating them for the increased risk, and perhaps restoring parity between markets. If the increase in enforcement were large enough, however, equilibrium might not be restored until all the dealers left that market.

The assumptions that dealers are identical and move around to balance the opportunities available are certainly artificial, but without some such assumptions the analysis would be hopelessly complex. Given the state of the data described in Chapter 2, it is unrealistic to expect to be able to calibrate more detailed models that, for example, subdivide the dealing population on the basis of experience.

These assumptions are similar in spirit to those commonly made in microeconomics. The dealers are analogous to firms and the markets to industries. It is assumed in elementary economics that industries are made up of a large number of identical firms and that free entry and exit ensure zero long-run profits. No one believes that those assumptions accurately model the business world, but at least some people believe that microeconomic theories based on them help explain phenomena in the real world.

If there are enough dealers that their number can reasonably be approximated as a continuous variable,[43] then the expected utility of dealing must be the same for all dealers. Call this common level of utility w_0.

The model developed below will explore the effects of cracking down on one market. If there are many markets in the city, then what happens in one will not appreciably affect the others, so this common level of utility will be treated as an exogenous constant. Since this level of utility is always available, if the utility in a market falls below w_0, dealers will leave that market. If it rises above w_0, dealers will enter. Thus w_0 is the reservation wage of the dealers.

It will further be assumed that all drug sales are identical, and that they yield a generalized profit π. By generalized profit it is meant the sale price minus the dealer's cost of doing business, including the costs imposed by other market participants and conventional police enforcement, but exclusive of the effects of the crackdown. Conventional enforcement includes enforcement by uniformed patrol officers not specifically directed to make narcotics arrests; crackdowns might be conducted by plain-clothes narcotics detectives and specially assigned uniformed patrols.

The assumption that π is constant bears some discussion. Within a city retail prices do not vary appreciably from market to market.[44] If they did, mobile customers would not patronize the expensive markets. Also, since dealers are envisioned as choosing a market each day, there is no reason to believe the dealers in one market are able to obtain consistently drugs at lower prices than dealers in other markets can. So the monetary profit per transaction probably does not vary from market to market within a city.

Non-monetary factors such as violence from other market participants might, however, vary from market to market. This possibility will be ignored below, but if this is not reasonable for a particular application, the equation should be re-interpreted accordingly.

The assumption that all transactions are identical also ignores differences between different kinds of drugs.[45] There are substantial differences between retail transactions for different

[43] Modeling the number of dealers in a market as a continuous variable can also be justified on the grounds that dealers can spend only part of the day in a particular market.

[44] Garreau (1989) notes that retail prices are uniform throughout Washington, D.C.

[45] Reuter and Kleiman (1986, pp.328-334) argue that local enforcement works best against heroin markets.

drugs, and there is evidence that dealers specialize and often sell only one kind of drug.[46] Nevertheless, it seems likely that explicitly identifying different markets for different drugs would add more notation and complexity than insight to the model, so differences between drugs and implications of polydrug use are not addressed. Depending on the application, however, such considerations could be important, and in those cases they should be kept in mind when interpreting the model. Let

N = the number of dealers in the market of interest and
Q = the number of sales per day[47] in that market.

Then each dealer in the market earns wage $\frac{\pi Q}{N}$. Now let

E = increment in enforcement pressure, above and beyond the baseline level, that is placed on the market during the crackdown.

Since dealers are assumed to share the burden of enforcement equally, each dealer is exposed to an enforcement effort of E/N. (The burden of conventional, "non-crackdown" enforcement is assumed to be a constant that is incorporated into the generalized net profit π.)

The definition of E is intentionally vague, in no small part because crackdown strategies vary from city to city depending on the nature of the problem.[48] It is some function of the probability of arrest, the likelihood an arrest will lead to a conviction, the likely punishment in the case of a conviction, and so on. Introducing a detailed model of the criminal justice system would distract attention from the characteristics of the market itself, which is the subject of interest. Treating enforcement pressure as a single exogenous variable is similar to microeconomists' treating the wage rate as a fixed, exogenous parameter.

Assuming that individual dealers suffer an enforcement related cost of E/N implicitly assumes that the total cost enforcement imposes on dealers, E, is not itself a function of N. It may be that the total cost imposed is an increasing function of N if police have to expend less effort apprehending a suspect when there are many

[46] Garreau, 1989.

[47] Specific units, such as sales per day, are used for clarity of exposition, but it should be clear that other units would be equally acceptable.

[48] Hayeslip, 1989.

suspects. Considering the extreme makes the point; if there were no dealers, then the total cost enforcement could impose on dealers would be zero no matter how many resources were allocated to the crackdown.

For several reasons, however, it is not unreasonable to assume that such effects will be minor. For one, if the limiting factor is court time or prison space, not police time, then the cost of enforcement per dealer would be essentially E/N. Second, if the police spend more time obtaining a warrant, doing paperwork, processing arrested individuals, and testifying in court than they do actually apprehending suspects, then reducing the number of dealers would not greatly increase the average total number of police-hours per arrest. Third, as the number of dealers declines, actually observing a deal may become more difficult, but other tactics, such as buy-busts, may not be greatly affected. So the police might be able to maintain their productivity by stressing tactics whose effectiveness are relatively insensitive to the number of dealers. Finally, although enforcement agents' jobs might get harder as the crackdown progresses because they have fewer targets, it might get easier because they have more potential informants (previously arrested dealers) and more cooperation from neighbors (who may be less intimidated by dealers once they see the number of dealers begin to decrease).

So for now it will be assumed that the enforcement pressure experienced by an individual is E/N, but Section 4.13 will explore how the results below would be affected if this were not a reasonable approximation.

Assume for now that the dealers' utility depends only on their wage and the enforcement pressure. Then one simple model of the flow of dealers in and out of a market is

$$\frac{dN}{dt} = c_1\left[U\left(\frac{\pi Q}{N}, \frac{E}{N}\right) - w_0\right], \qquad (4.1)$$

where U(x,y) is the utility a dealer derives from a wage of x when the enforcement pressure experienced is y.

Equation 4.1 suggests that if dealers in a particular market have a utility greater than the utility available elsewhere, more dealers will move to this market, and if their utility is smaller, some will exit. The constant c_1 governs how quickly this adjustment is made. If c_1 is large, then dealers change markets quickly. If c_1 is small, then differences in utility between markets could persist for some time.

Some assumption must be made about the functional form of U(x,y) for the analysis to proceed. It is clear that U(x,y) should be increasing in x and decreasing in y. Further, since the first argument is a measure of profits and the second of cost (risk), it seems reasonable that U(x,y) should have the form U(x,y) = f(x) − g(y).

The utility of income is generally modelled as being concave, but approximated as linear for small ranges. The balloon model assumes that a linear approximation is adequate, so U(x,y) = x − g(y).

For people who are risk neutral, risks can be summarized by taking the expected cost of the corresponding risk. In that case, g(y) would be linear in y. One might think dealers are risk neutral, perhaps even risk seeking. After all, they have selected a very risky profession. But most people, probably even dealers, are at least somewhat risk averse, suggesting that g"(y) > 0. Analytically the most convenient increasing, convex function is $g(y) = y^{\gamma}$ for $\gamma \geq 1$, so that function will be used. Initially, however, for expository purposes, only the risk neutral case of $\gamma = 1$ will be considered. Section 4.6 generalizes the results to all $\gamma \geq 1$.

If U(x,y) = x − y then Equation 4.1 becomes

$$\frac{dN}{dt} = c_1\left[\frac{\pi Q}{N} - \frac{E}{N} - w_0\right]. \qquad (4.2)$$

Ideally one would solve Equation 4.2 to find the number of dealers N as a function of the enforcement pressure E and time. For several reasons, however, the analysis in this chapter focuses on the steady state solution obtained by setting Equation 4.2 equal to zero instead of the dynamic solution of N as a function of time. First, the steady state analysis is all that is needed to derive many useful insights. Second, one can feel a great deal more confidence in the statement that

$$\mathrm{Sgn}\left(\frac{dN}{dt}\right) = \mathrm{Sgn}\left(\frac{\pi Q}{N} - \frac{E}{N} - w_0\right) \qquad (4.2a)$$

than one can in the exact form of Equation 4.2, and Equation 4.2a is all that is needed for the steady state analysis. Third, even if the form of Equation 4.2 were correct, there is little hope of measuring the parameter c_1.

One must be careful then in interpreting statements about how characteristics of the market, such as the number of dealers, are related to the enforcement pressure E. For example, it will be determined that the steady state number of dealers is decreasing in E. This should be interpreted to mean that if the market is in steady state with a particular level of enforcement E_1, and if enforcement is subsequently increased to a new level E_2, then after the market has returned to equilibrium, there will be fewer dealers than before. Furthermore, if the new level of enforcement had been E_3 not E_2, and E_3 is greater than E_2, then the new equilibrium number of dealers would have been even smaller. One should not think of the level of enforcement E as steadily increasing, unless it changes slowly enough that a quasi-static equilibrium, of the sort assumed in elementary thermodynamics, is maintained.

So the objective is to relate the steady state characteristics of the market to the (constant) level of enforcement E. Even this cannot be done yet because, although π, c_1, and w_0 are constants, the number of sales Q almost certainly depends on the number of dealers N. It might also depend on enforcement against dealers (E) directly, but this possibility will be ignored.

This assumption is significant because some crackdowns do explicitly seek to arrest users. However, the arrest risk for users, even during a crackdown, is quite low. Also, Section 4.11 does examine how effort against dealers should be balanced against efforts to control demand.

How exactly does the number of sales depend on the number of dealers? First, if there were no dealers, there would be no dealing ($Q(0) = 0$). Second, increasing the number of dealers would probably never reduce the number of sales ($Q'(N) \geq 0$), and it would probably increase sales at a decreasing rate ($Q''(N) < 0$).

This last comment need not hold. It may be that the presence of many dealers creates a sense of social acceptability or peer pressure that may induce customers to buy more. In that case sales could increase more than proportionately in the number of dealers.

This possibility will be ignored, however, on the principle of diminishing returns. Consider the volume of sales to be the product. Demand and dealers are the inputs. Then if conventional economic wisdom carries over to this case, for a fixed level of demand, increasing the number of dealers would probably increase the number of sales, but at a decreasing rate.

Beyond these observations though, it is difficult to say much for certain about Q(N).

At first one might think that Q(N) would be almost linear in N. That would fit the old-fashioned view that dealers are "pushers" who create their demand. But that view has been largely rejected.[49]

In many markets most of the customers drive in from outside the neighborhood. That might lead one to think Q(N) is almost independent of N, because, as long as there were one or two dealers present, almost everyone who comes to the market could make their purchase.

Actually it might take more than one or two dealers. Retailers frequently do not keep drugs in their possession. When they identify a potential customer they return to their "cache" to get the drugs. So one sale can keep a retailer busy for several minutes, even if the transaction itself is brief.

More importantly, for several reasons mobile customers are more likely to go to markets with lots of dealers. First of all, the more dealers there are, the more likely they are to score quickly, and customers have an incentive not to spend any more time in the market than is absolutely necessary.[50] Second, when there are many dealers, competition might lead them to give better service, and perhaps even to give discounts. Finally, there is safety in numbers. If the market is large, then even if the police decide to arrest a user there is less chance that any particular user will get caught.

The discussion below considers two extreme forms for Q(N) first and then a more plausible intermediate case. Section 4.5 solves these three special cases as a prelude to Section 4.6, which solves a more general version of the model. The properties of the more general solution are foreshadowed by those of the three specific forms, so the solutions to the specific forms are discussed at length.

The first extreme is a "seller's market" in which the volume of sales is limited only by the dealers' selling capacity. The second extreme is a "buyer's market" in which demand is fixed and dealers compete for the chance to make sales. In the intermediate case the number of sales Q(N) is an increasing but strictly concave function, so total sales increase when the number of dealers increases, but the number of sales each dealer makes decreases.

[49] See, for example, Kaplan, 1983a, pp.25-32.
[50] Garreau, 1989.

Before describing the solutions, it may be useful to briefly review the key assumptions that have been made.

(A1) Dealers are identical and interchangeable.
(A2) Dealers go to the market offering the greatest return.
(A3) The number of dealers can be modeled as a continuous variable.
(A4) Cracking down on one of many markets in a city does not significantly influence dealing in the other markets.
(A5) All drug sales yield the same generalized profit π.
(A6) All dealers experience a crackdown pressure of E/N.
(A7) Dealers' utility as a function of their expected (generalized) profit x and individual enforcement pressure y is $U(x,y) = x - y^{\gamma}$.
(A8) Dealers enter the market if and only if U(x,y) is greater than the dealers' reservation wage w_0.
(A9) Sales are a function only of the number of dealers, N, and are not a function of the enforcement pressure E directly.

4.5 Solutions for Three Special Cases

4.5.1 A Dealer's Market ($\gamma = 1$ and $\beta = 1$)

Consider first the extreme case in which there is so much demand that all dealers sell the maximum (q_{max}) they can, i.e.

$$Q(N) = q_{max} N. \tag{4.3}$$

In such a market availability is severely limited relative to demand. It might characterize a market the day after most of the dealers were arrested. The remaining ones, were they brave or foolish enough to deal, would be able to sell essentially as much as they could obtain. Another plausible scenario would be that the wholesalers supplying most of the street-level dealers have been arrested, so only a fraction of the dealers usually operating in the market are able to deal, and each of them can only obtain enough drugs to make q_{max} sales.

A market in which dealers can sell as much as they want would probably attract other dealers unless, perhaps, the police pressure per dealer were quite high. In that case some dealers would exit. The revenues of the remaining dealers would remain the same, but their costs would increase, so still more would exit. Either way one would not expect the market to be stable. The model confirms this intuition.

If $Q(N) = q_{max} N$ then

$$\frac{dN}{dt} = c_1\left[\pi q_{max} - \frac{E}{N} - w_0\right]. \tag{4.4}$$

Suppose there were an equilibrium with enforcement $E = E_0$. Setting Equation 4.4 equal to zero shows this implies that

$$N = \frac{E_0}{\pi q_{max} - w_0}. \tag{4.5}$$

If E increases slightly, $\frac{dN}{dt}$ becomes negative. Then as N decreases, $\frac{dN}{dt}$ becomes more negative until eventually N goes to zero. On the other hand, if for some reason N were to increase slightly, $\frac{dN}{dt}$ would become positive. As N grew, $\frac{dN}{dt}$ would become more positive, and N would increase without bound (or until $Q = q_{max} N$ no longer held).

This suggests several things. First of all, it is unlikely that Q(N) is linear (or convex) for values of N that are actually observed. Second, it suggests that markets that have been impacted by enforcement, either by reducing the number of retailers or by cutting off most of their supply, may be vulnerable. Either of these might make Q(N) nearly linear. Then if enforcement becomes sufficiently strict, the market will collapse. In particular, once the expected utility falls below w_0, dealers begin to exit. As dealers exit, the enforcement pressure per dealer increases while their monetary profit stays the same, so more dealers exit, making the remaining dealers still worse off. This positive feedback feeds on itself until all the dealers have moved elsewhere.

Furthermore, once the dealers have all left, a small amount of enforcement ($E > \pi q_{max} - w_0$) will keep any individual from beginning to deal. That is because if one person starts dealing, he or she would suffer all the enforcement pressure. Thus a low level of dealing (in this case no dealing) is a stable equilibrium even though there is relatively little enforcement. No individual dealer has an incentive to deviate from his or her current strategy.

However, if $N > \frac{E}{\pi q_{max} - w_0}$ dealers agreed to begin dealing at once, they could "jump start" the market by spreading the enforcement costs among them. Once established, the market would grow without bound unless enforcement pressure were subsequently increased.

To illustrate this more clearly, consider the case in which there are just two potential dealers, each deciding repeatedly but independently whether or not to deal. Figure 4.1 shows the payoff matrix.

Figure 4.1:
Payoff Matrix for Two Dealers

$$\begin{array}{lcc} & \text{Deal} & \text{Do Not Deal} \\ \text{Deal} & \left(\pi q_{max} - \frac{E}{2}, \pi q_{max} - \frac{E}{2}\right) & \left(\pi q_{max} - E, w_0\right) \\ \text{Do Not Deal} & \left(w_0, \pi q_{max} - E\right) & \left(w_0, w_0\right) \end{array}$$

Suppose

$$\pi q_{max} - w_0 < E < 2\left(\pi q_{max} - w_0\right) \tag{4.6}$$

so

$$\pi q_{max} - \frac{E}{2} > w_0 > \pi q_{max} - E. \tag{4.7}$$

There are two stable equilibria: one in which neither person deals and one in which both deal. If initially both are dealing, they will continue to do so because their payoffs are the largest possible. Suppose instead that initially neither is dealing. They would prefer that they were both dealing, and if they could agree to commit to deal in the next period they would. If they cannot collude and bind themselves to that course of action, however, then the threat of incurring the full weight of the enforcement pressure (i.e. receiving a payoff of $\pi q_{max} - E$) might deter them from beginning to deal.

There is a lesson here for the role gangs might play in starting drug markets. If every dealer acts independently and no collusion is possible, an empty market is stable. But if dealers could coordinate, they could improve their lot by all starting to deal. Gangs might provide such a coordinating mechanism.

This model suggests several other important lessons. First, there cannot be a stable equilibrium in a region where Q(N) increases linearly in N. By the same reasoning, Q(N) cannot be convex at a stable equilibrium.

Second, there may be a threshold level of enforcement beyond which a positive feedback effect is created. Once this feedback effect takes hold, simply continuing the same level of enforcement pressure may wipe out the market. Hence the benefits of enforcement may be a highly nonlinear function of the enforcement pressure applied.

Finally, it may take considerably less effort to prevent a market from springing back than was required to make it collapse in the first place. However, more effort is needed if gangs or some other coordination mechanism exists.

As will be seen below, these observations are not artifacts of the extreme assumption that Q(N) is linear in N.

4.5.2 A Buyer's Market ($\gamma = 1$ and $\beta = 0$)

Consider next the other extreme: a market in which the number of sales Q is a constant Q_0 independent of the number of dealers. Such a market is saturated with dealers. If another dealer arrives, the number of sales does not increase; there are just more dealers fighting over the fixed number of sales.

If Q(N) is a constant Q_0 then

$$\frac{dN}{dt} = c_1\left[\frac{\pi Q_0}{N} - \frac{E}{N} - w_0\right], \tag{4.8}$$

and the equilibrium number of dealers is just

$$N = \frac{1}{w_0}\left[\pi Q_0 - E\right]. \tag{4.9}$$

Hence

$$\frac{dN}{dE} = \frac{-1}{w_0} \tag{4.10}$$

which is constant. There is no positive feedback effect. Reducing the number of dealers increases the enforcement cost per dealer, but it also increases their profits. No matter how much enforcement pressure is applied, it takes the same amount of additional pressure to achieve an incremental reduction in the number of dealers. There is no threshold number of dealers below which the market collapses.

Furthermore, if all markets in the region have constant demand, then since w_0 has one fixed value throughout the region, it makes no difference how enforcement is allocated between markets within the region. No matter how the extra enforcement pressure is distributed, the number of dealers will be reduced by

$$\Delta N = -\frac{\Delta E}{w_0}. \qquad (4.11)$$

Although in a saturated market reducing the size of the market as measured by the number of dealers is difficult, when the volume of sales is used as a measure of the size of the market the picture is even more bleak; by assumption the number of sales is a constant independent of the number of dealers.

Note also that even after the market has been eliminated, the police would have to maintain all the enforcement pressure needed to collapse the market just to keep it from springing back.

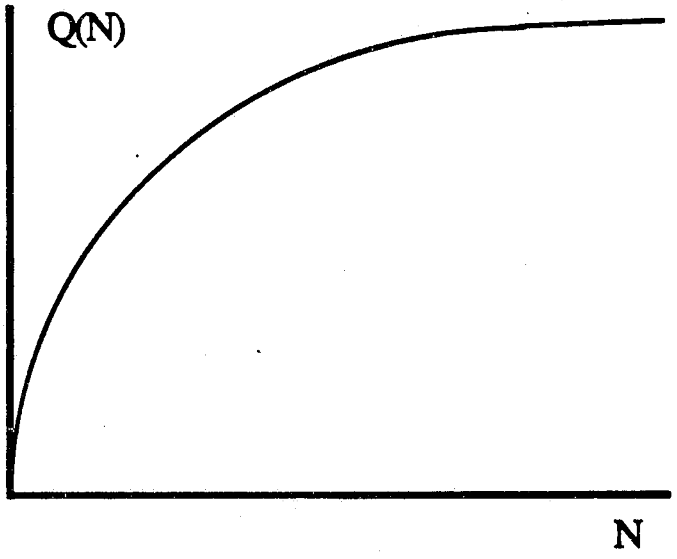
Of course it is unrealistic to think that the sales potential for a start-up market is as great as it is for an established market, so this Q_0 would not have the same numerical value as would be appropriate for Equation 4.8. But it makes the point that neighborhoods with a fixed sales potential may require a substantial amount of enforcement pressure to prevent their becoming a drug market even if there is no dealing there presently. Streets that buyers travel to get to established markets may be a prime example of such neighborhoods.

The case when Q(N) is constant yields other important lessons. First, if demand is not responsive to the number of dealers then targeting dealers may not be an effective strategy. Second, if a market is adequately described by this model with Q(N) constant, it is easy to determine whether the crackdown will be able to eliminate all the dealing because progress is linear in effort. If when half the effort available has been applied, more than half the dealers remain active, then even when the crackdown is fully implemented, it will not be able to eliminate all of the dealing. Finally, in a market with fixed demand, intensive crackdowns are not productive. The market will always spring back unless the level of enforcement needed to clean up the market in the first place is maintained even after the dealers have been driven away. Hence when Q(N) is constant, the model suggests dramatically different possibilities for positive feedback and preventing a market from springing back than was the case with Q(N) linear in N. The next form for Q(N) considered is an intermediate and probably more realistic case.

4.5.3 An Intermediate Case ($\gamma = 1$ and $\beta = 1/2$)

The two cases considered so far are extremes. It is hard to imagine measuring Q(N), but it seems reasonable that it is a concave function such as the one depicted in Figure 4.2, not linear or constant.

Figure 4.2:
Sales as a Function of the Number of Dealers, Q(N)



If there are many dealers in the market, adding a few more is unlikely to generate many additional sales, so for large N, Q(N) may be nearly constant. Likewise, when N is very small there could be so many potential customers relative to the number of dealers that each dealer sells as much as he or she can obtain. So for small N, Q(N) may be approximately linear.

If so, there must be some transition for intermediate values of N. The author certainly does not know what this transition is, but choosing

$$Q(N) = \alpha N^{\beta} \qquad \beta \in [0,1] \qquad (4.12)$$

seems like a reasonable guess. It has the desired shape, and it is analytically convenient. The solutions above correspond to $\beta = 1$ and

0, respectively. This subsection considers the intermediate case when $\beta = {}^1/_2$. Section 4.6 gives the solution for arbitrary $\beta \in (0,1)$.

With $\beta = {}^1/_2$

$$\frac{dN}{dt} = c_1 \left[\frac{\pi \alpha \sqrt{N}}{N} - \frac{E}{N} - w_0 \right]. \tag{4.13}$$

Setting this equal to zero to obtain the steady state solution implies that $\pi \alpha \sqrt{N} - E - \frac{w_0}{N} = 0$. This equation is quadratic in $\sqrt{N}$. Its roots are

$$\sqrt{N} = \frac{\pi \alpha \pm \sqrt{(\pi \alpha)^2 - 4 w_0 E}}{2 w_0}. \tag{4.14}$$

The larger root gives the stable equilibrium, so

$$N(E) = \frac{(\pi \alpha)^2 - 2 w_0 E + \pi \alpha \sqrt{(\pi \alpha)^2 - 4 w_0 E}}{2 w_0^2}. \tag{4.15}$$

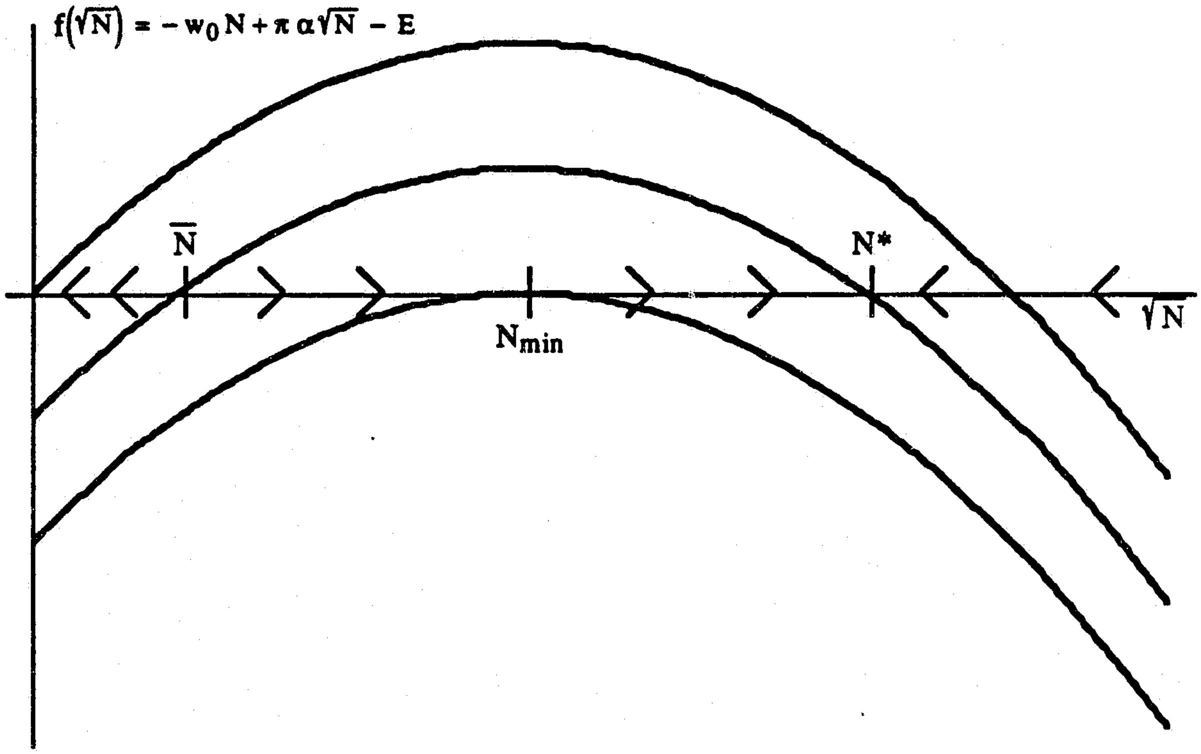
Figure 4.3 shows visually how the number of dealers N(E) is affected by changes in the enforcement pressure. It plots the function

$$f(\sqrt{N}) = - w_0 N + \pi \alpha \sqrt{N} - E \tag{4.16}$$

which has the same sign as $\frac{dN}{dt}$.

Figure 4.3:

The Sign of $\frac{dN}{dt}$ for Various Levels of Enforcement



The highest curve corresponds to no special enforcement, E = 0. The roots give the two values of $\sqrt{N}$ that satisfy Equation 4.15, namely $\sqrt{N} = 0$ and $\sqrt{N} = \frac{\pi \alpha}{w_0}$. Since the area is presumed to be a market, the first root can be ignored. So the number of dealers when there is no enforcement is

$$N_{max} \equiv N(E = 0) = \left(\frac{\pi \alpha}{w_0}\right)^2. \tag{4.17}$$

This quantity is denoted N_{max} because it is the largest number of dealers the market will ever support.

One can similarly define

$$Q_{max} \equiv Q(E = 0) = \alpha \, N_{max}{}^{1/2} = \frac{\pi \alpha^2}{w_0}. \tag{4.18}$$

This quantity is denoted Q_{max} because it is the maximum sales volume the market can support.

The middle curve in Figure 4.3 shows the situation with a moderate level of enforcement. The right hand root, labeled N^*, gives the stable equilibrium. If $N > N^*$, $\frac{dN}{dt} < 0$. If N is less than N^* but greater than the left hand root $\tilde{N}$, then $\frac{dN}{dt} > 0$ so the number of dealers increases to N^*. If, on the other hand, $N < \tilde{N}$, then $\frac{dN}{dt} < 0$ and the market collapses.

This makes sense. If there are "too many" dealers, the number of customers per dealer is small, so dealers exit. If there are "too few" dealers they incur the full weight of enforcement and are driven off. If there are an intermediate number of dealers, each prospers and more dealers enter.

At higher enforcement levels, the intercept of $f(\sqrt{N})$ decreases, which shifts the equilibrium number of dealers N^* to the left. As the level of enforcement approaches that depicted in the bottom curve in Figure 4.3, the region where $\frac{dN}{dt}$ is positive shrinks. When

$$E = \frac{(\Pi\,\alpha)^2}{4\,w_0} \tag{4.19}$$

there is just a single root, an unstable equilibrium. Any additional enforcement pressure will make $\frac{dN}{dt}$ negative for all N. As N decreases, $\frac{dN}{dt}$ becomes more negative and the market collapses.

Hence one can visualize enforcement as pressing the curve in Figure 4.3 down. As the intercept decreases, the stable equilibrium number of dealers decreases until the roots given by Equation 4.14 become complex. At that point the market collapses.

The enforcement level needed to make the market collapse will be called E_{max} because it is the maximum level of enforcement the market will ever experience in steady state. The notation is somewhat confusing because E_{max} is the minimum level of enforcement needed to collapse the market. It is easy to remember what E_{max} means, however, by remembering the story of gradually increasing enforcement (maintaining quasi-static equilibrium) until the market collapses. The amount of enforcement exerted just before the market collapses is E_{max}.

Setting the radical equal to zero gives

$$E_{max} = \frac{(\Pi\,\alpha)^2}{4\,w_0}. \tag{4.20}$$

When enforcement is at its maximum level, the market is at its minimum steady state size. So, the equilibrium number of dealers just before the market collapses will be called N_{min},

$$N_{min} \equiv N(E_{max}) = \left(\frac{\pi\,\alpha}{2\,w_0}\right)^2 = \frac{N_{max}}{4}. \tag{4.21}$$

The market collapses when there are a quarter as many dealers as there would be when there is no enforcement.

Likewise one can define Q_{min} to be the amount of dealing when the market is at its minimum size, just before it bursts.

$$Q_{min} \equiv Q(E_{max}) = \frac{\pi\,\alpha^2}{w_0} = \frac{Q_{max}}{2}. \tag{4.22}$$

Figure 4.3 suggests that there is a positive feedback effect. One can see visually that when $E \approx 0$ increasing E a little will not reduce the equilibrium number of dealers N* very much. But as the level of enforcement approaches that depicted in the bottom curve in Figure 4.3, small increases in E appreciably reduce N*.

Taking derivatives of N(E) with respect to E confirms this.

$$\frac{dN(E)}{dE} = \frac{-1}{w_0}\left(1 + \frac{\pi\,\alpha}{\sqrt{(\pi\,\alpha)^2 - 4\,w_0\,E}}\right) < 0 \tag{4.23}$$

and

$$\frac{d^2N(E)}{dE^2} = \frac{-2\,\pi\,\alpha}{\left((\pi\,\alpha)^2 - 4\,w_0\,E\right)^{3/2}} < 0. \tag{4.24}$$

Since $\frac{d^2N(E)}{dE^2} < 0$ there is a positive feedback effect; the marginal effectiveness of a unit of enforcement increases with the level of enforcement.

Hence, as was the case above with $Q(N) = q_{max}N$ but not with $Q(N) = Q_0$, if $Q(N) = \alpha N^{1/2}$, one can accomplish more by focusing on one market than by spreading resources over many markets.

Equation 4.15 is fairly complex so it is hard to see intuitively how N depends on E. Normalizing quantities helps, so define

$$n \equiv \frac{N}{N_{max}},\ q \equiv \frac{Q}{Q_{max}},\ \text{and}\ e_N \equiv \frac{E}{E_{max}}. \qquad (4.25)$$

The notation e_N is used instead of e to avoid confusion with the constant $e \approx 2.71828$. The subscript N is used to denote "normalized".

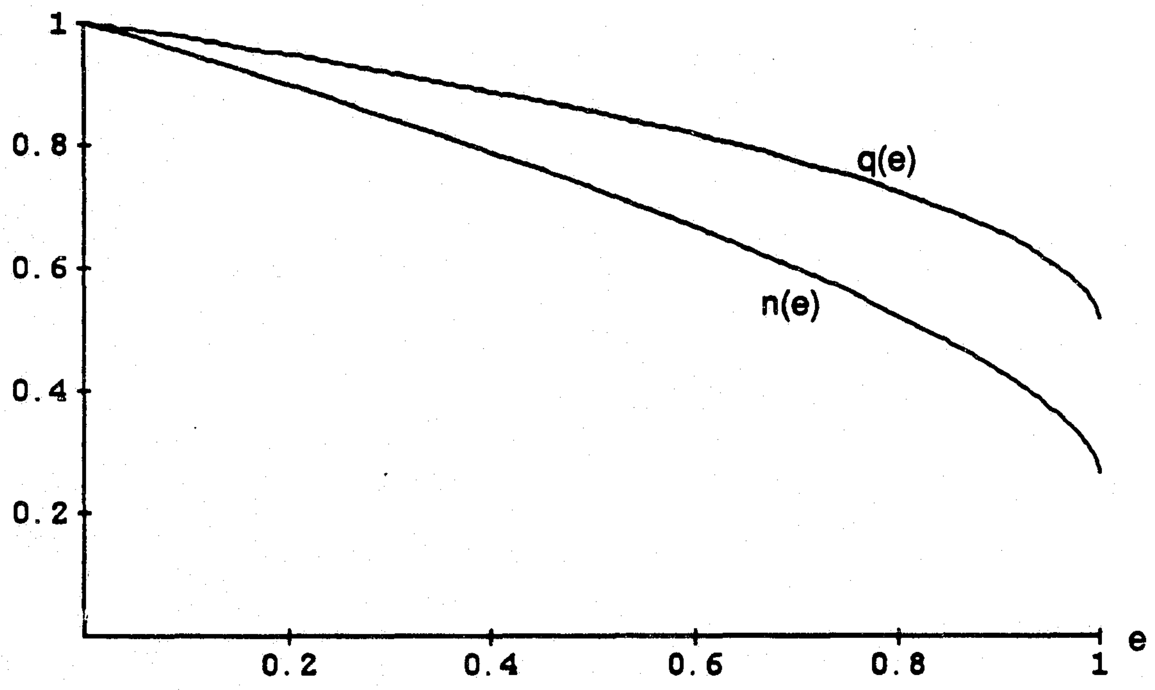
With these definitions one can show that

$$n = \left(\frac{1 + \sqrt{1 - e_N}}{2}\right)^2 \ \text{and} \qquad (4.26a)$$

$$q = \frac{1 + \sqrt{1 - e_N}}{2}. \qquad (4.26b)$$

Figure 4.4 plots Equations 4.26a and 4.26b. They show clearly that the marginal efficacy of an extra unit of enforcement increases with the enforcement level and that when the market has been reduced to $^1/_4$ or $^1/_2$ of its original size (depending on whether one measures the number of dealers or the amount of dealing) the market collapses.

Figure 4.4:
The Number of Dealers and the Level of Enforcement
(β = 1/2 and γ = 1)



Inverting Equation 4.26a gives the level of enforcement needed to maintain equilibrium for any given number of dealers n. Hence, to prevent the market from springing back when

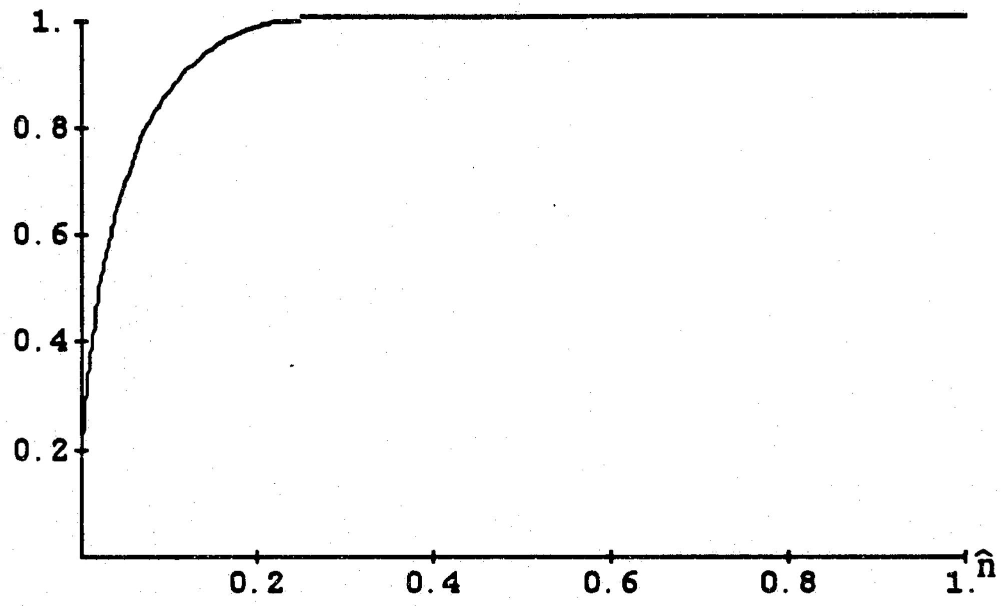
$$\hat{N} \equiv \hat{n}\,N_{max} \tag{4.27}$$

dealers try to start dealing, one needs to maintain a level or pressure equal to

$$\hat{E} = \begin{cases} 4\left(\sqrt{\hat{n}} - \hat{n}\right)E_{max} & \text{if } \hat{n} \le \frac{1}{4} \\ E_{max} & \hat{n} \ge \frac{1}{4} \end{cases} \tag{4.28}$$

As Figure 4.5 shows, this suggests that unless a high level of enforcement pressure is maintained after the market is shut down, even a relatively small group of dealers could successfully "jump start" the market.

Figure 4.5:
Effort Needed to Keep the Market from Springing Back
(β = 1/2 and γ = 1)



Hence the intermediate case of $Q(N) = \alpha N^{1/2}$ displays some of the characteristics of each of the two extreme cases. There is a positive feedback effect, so focusing enforcement pressure may accomplish more than spreading it uniformly across all markets. And the authorities should not necessarily conclude that there is no hope for enforcement-oriented solutions just because a modest amount of pressure seems to accomplish little. It may be that doubling the effort will more than double the results.

On the other hand, if demand is of the form $Q(N) = \alpha N^{1/2}$, it may be relatively easy for a small group of dealers to "jump start" the market. So after enforcement pressure shuts down a market, if even 10 - 20% of the displaced dealers coordinate and begin dealing again, they may be able to resurrect the market.

Ideally one would next solve for the general case of $Q(N) = \alpha N^{\beta}$ for $\beta \in [0,1]$. Parameterizing the answer by β might show how the phenomena discussed above, such as the positive feedback effect and the ability of a market to spring back, depend on β. Unfortunately simple analytic solutions for N(E) do not exist for all values of β.

One can obtain a closed-form solution when $\beta = {}^1/_4$ and $\beta = {}^3/_4$ by solving a quartic equation, but the algebra is intimidating. Section 4.7 finds N(E) when $\beta = {}^1/_3$ and $\beta = {}^2/_3$ by solving a cubic equation. "Interpolating" between the solutions for $\beta = 0, {}^1/_3, {}^1/_2, {}^2/_3$, and 1 probably gives adequate intuition about intermediate values of β.

The next section examines the general case in which β takes on any value between 0 and 1 and γ can take on values greater than unity. Fortunately, even though one cannot obtain a closed form solution for N(E) for arbitrary γ and β, one can learn a good deal about those solutions indirectly.

4.6 More General Results

Section 4.5 solved the balloon model for demand parameter values $\beta =$ 0, 1/2, and 1 when dealers were assumed to be risk neutral ($\gamma =$ 1). This section extends those results by allowing the demand parameter β to be any value between zero and one, and the risk aversion parameter to take on values other than one. In general one cannot obtain explicit solutions of the steady state market size as a function of enforcement pressure (i.e. for N(E) and Q(E)), but one can find N_{max}, Q_{max}, E_{max}, N_{min}, Q_{min}, and E(N).

With arbitrary β and γ Equation 4.2 becomes

$$\frac{dN}{dt} = c_1\left[\frac{\pi\,\alpha\,N^\beta}{N} - \left(\frac{E}{N}\right)^\gamma - w_0\right] = F(N). \tag{4.29}$$

Figure 4.6 shows the general shape of F(N) when there is no crackdown, i.e. of $\frac{dN}{dt} = c_1\left(\pi\,\alpha\,N^{\beta-1} - w_0\right)$, for $\beta \in (0,1)$.

Figure 4.7 does the same for $-c_1\left(\frac{E}{N}\right)^\gamma$, the contribution enforcement makes to $\frac{dN}{dt}$.

Figure 4.6:

$\frac{dN}{dt}$ With No Enforcement

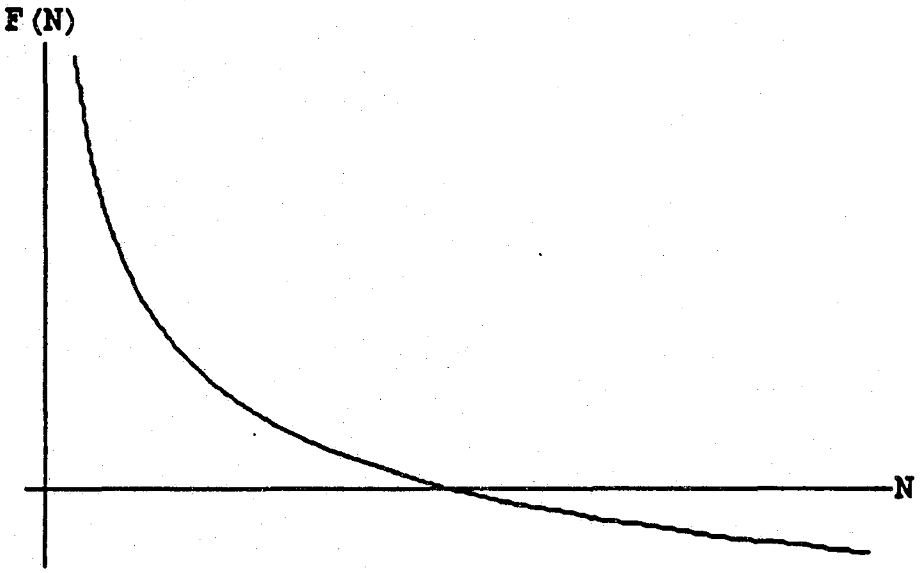
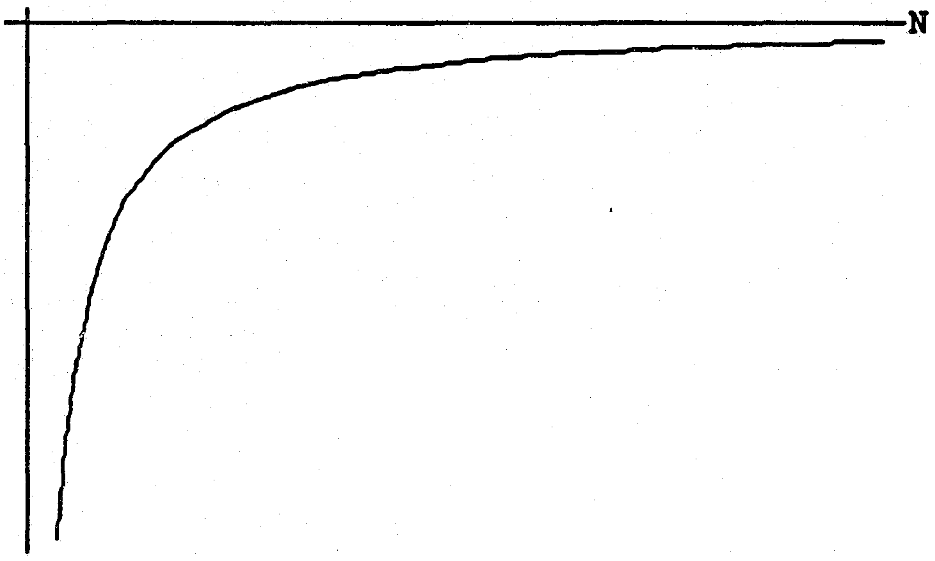


Figure 4.7:

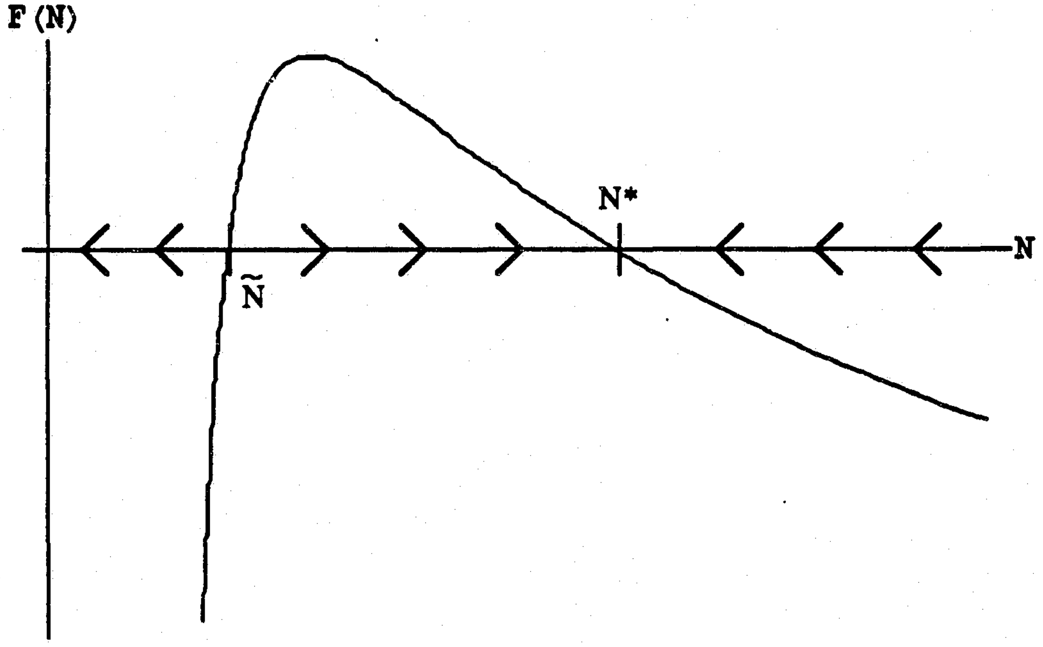
Enforcement's Contribution to $\frac{dN}{dt}$



Since γ is greater than one, $\gamma > 1 - \beta$. So $\left(\frac{1}{N}\right)^{\gamma} >> N^{\beta-1}$ for $N \approx 0$. Hence F(N) has the shape shown in Figure 4.8

Figure 4.8:

$\frac{dN}{dt}$ As a Function of N



One can also show algebraically that when $\frac{dN}{dt}$ is viewed as a function of N for N > 0, it has two roots and one maximum, asymptotically approaches $-w_0$ for N large, and goes off to minus infinity as N approaches zero. Setting the derivative of F(N) with respect to N equal to zero shows that the maximum occurs at

$$\bar{N} = \left[\frac{\gamma\, E^{\gamma}}{\pi\, \alpha\, (1 - \beta)}\right]^{1/(\gamma+\beta-1)}. \qquad (4.30)$$

Since

$$\frac{d^2F(\bar{N})}{dN^2} < 0, \qquad (4.31)$$

it is indeed a maximum.

As the arrows on the horizontal axis of Figure 4.8 suggest, for a given level of enforcement E, if $N < \tilde{N}$ the market will collapse, but if $N > \tilde{N}$ the number of dealers will converge to N*.

As was the Case with Figure 4.3, an enforcement campaign can be thought of as pulling down the curve in Figure 4.8. It does this by making the enforcement component of $\frac{dN}{dt}$, depicted in Figure 4.7, larger in absolute value. If the enforcement pressure is sufficient to pull the entire curve below the horizontal axis, the market will collapse. "Collapse" is an appropriate word because initially the number of dealers will decrease rapidly, but the rate of decrease will slow for $N = \bar{N}$. Then once N drops below $\bar{N}$, the rate of decrease will increase; i.e., the market will collapse.

As was discussed above, after the market has collapsed, some of the enforcement pressure can be removed without allowing any dealers to come back. In particular, once a market has been cleaned up, attempts by less than $\tilde{N}$ dealers to "jump start" the market would fail. If more than $\tilde{N}$ dealers arrive and start to deal, they could overwhelm the enforcement present and the market would grow to N*. This will be discussed in greater detail in Section 4.8.

Consider next what happens if a crackdown is begun in which the enforcement effort is less than E_{max}. That would pull the curve in Figure 4.8 down, shifting N* to the left. So the number of dealers in the market would decrease until it reached its new steady state value.

When the crackdown ends, however, N* will equal N_{max} again and the number of dealers will grow back to its original value. This suggests that the balloon metaphor is a good one. Pressing down a little on a balloon (market) accomplishes nothing because as soon as the extra pressure is removed, it springs back to its original volume. However, if enough pressure is applied, the balloon (market) will pop (collapse). Then even if the pressure is removed the balloon (market) will not spontaneously inflate again (the market will not spring back).

Hence an important lesson of the balloon model is that crackdowns should only be undertaken if there are sufficient resources to collapse the market. Simply denting a market accomplishes nothing in the long run.

A key implication of this is that police should only crackdown on one market at a time. Far more is accomplished by collapsing one market than by denting two.

Most of this chapter focuses on steady state results, but this is one point where it is important to think of the dynamics. There are two ways that a crackdown could only dent and not collapse a market. The first is simply that the enforcement pressure is not great enough, i.e. $E < E_{max}$. Then no matter how long the crackdown remained in place, it would never collapse the market.

The second way is by imposing a crackdown with $E > E_{max}$ but not leaving it in place long enough. If E is very large, $\frac{dN}{dt}$ will be negative. But if E returns to a smaller value before N decreases below the value of N_{min} corresponding to the sustained level of enforcement, the market will not collapse. Short, intense, headline-grabbing crackdowns may be a waste of resources. The crackdown must remain in place long enough to drive dealers away.

When thinking about the model this point may seem transparent, but short, sharp crackdowns have in fact been implemented. For example, the Hartford, Connecticut police arrested 3,666 persons for narcotics violations in 1988,[51] on average about 10 a day. Then on one day in May, 1989, they mounted Operation Pointed Eagle in which they arrested 171 people on narcotics violations in one day, but arrests soon returned to their usual level.

One can imagine how the dealers in Hartford who were not arrested might have reacted to the Pointed Eagle Operation. They would very likely have stayed home the next day because the "heat was on."[52] And maybe the next day too. But soon they would realize the risk of arrest was no higher than it was before the Pointed Eagle Operation, and many would resume dealing.

Hartford is not alone. The Florida Sheriff's Department arrested 2,224 people on June 30 and July 1 in a crackdown on crack dealing.[53] In two subsequent intense operations they made over 4,000 more arrests.[54]

Just because crackdowns are focused geographically does not necessarily mean they should be concentrated in time as well. To illustrate this, consider a market with 100 dealers. One way to wipe out the market is to arrest all 100 dealers today. Another way is to arrest 25 today, scaring 50 away, and then arrest the remaining 25 tomorrow.

[51] Jetmore, 1989.

[52] Woodley, 1971, describes instances in which arrests made dealers cautious.

[53] Navarro, 1989.

[54] Navarro, 1990.

Building up pressure slowly and scaring dealers away instead of trying to arrest them all offers two major advantages. First, it requires imprisoning fewer people. Second, in as much as there is heterogeneity among dealers, it is more likely to punish the "hardest" criminals. If only some of the dealers who escape the first round of arrests try to deal on the second day, it is probably fair to assume they are in some sense the dealers who are least likely to reform. In effect, the second strategy allows for the analog of third-degree price discrimination.[55]

Thinking about the dynamics suggests another reason why it might make sense to attack one market at a time. This reasoning is only a conjecture, however, because it stretches the limits of the model's applicability. Suppose a city had enough resources to crack down on two markets with enough pressure to collapse both, but not much to spare. Then if it cracked down on both it might take a long time for the markets to shrink to the point where they collapse and disappear. In contrast, if all the pressure were placed on one market, it might collapse quite quickly. It might take less time (and hence less resources overall) to collapse the markets one at a time, than to attack them simultaneously.

In summary, the balloon metaphor's key characteristics hold for the balloon model with arbitrary values of the demand parameter β and the risk aversion parameter γ.

Unfortunately one cannot find the roots of F(N;E) and hence $\tilde{N}(E)$ and N*(E) for arbitrary values of β and γ. When γ and β are related in certain ways, expressions happen to simplify so closed form solutions exist. Section 4.7 presents the solutions for $\gamma = (1 - \beta)$, $\gamma = 3(1 - \beta)/2$, $\gamma = 2(1 - \beta)$, and $\gamma = 3(1 - \beta)$. There is no physical explanation for the special properties of these combinations of β and γ; they are mathematical artifacts.

More importantly, one can solve explicitly for N_{max}, Q_{max}, E_{max}, N_{min}, and Q_{min} for arbitrary β and γ. This is done next.

4.6.1 The Size of the Market Before a Crackdown

The value of N_{max} is simply that value of N for which $\frac{dN}{dt} = 0$ when E = 0. That value is

$$N_{max} \equiv N(E=0) = \left(\frac{\pi\alpha}{w_0}\right)^{1/(1-\beta)}. \qquad (4.32)$$

[55] Tirole (1988, Chapter 3) describes third degree price discrimination.

Not surprisingly the size of the market in the absence of a crackdown is positively related to the profitability of an individual transaction and to the proportionality constant of demand. Likewise it is negatively related to the value of dealers' reservation wage.

What is somewhat more surprising is the prominent role the demand parameter β plays in this expression. If $\beta = 0$ then doubling the profit per transaction will double the number of dealers. For $\beta > 0$, however, doubling the profit per transaction will more than double the number of dealers, and for $\beta \approx 1$, the number of dealers is very sensitive to the profit margin.

The explanation for this is the following. Increasing π increases the dealers' wage, so more dealers enter. Unless $\beta = 0$ this brings in more customers, which further increases the earnings. Then more dealers enter, bringing in more customers, until eventually a new equilibrium is reached. If each new dealer brings in many new customers ($\beta \approx 1$), the new market equilibrium will be considerably larger than it was previously.

The maximum volume of sales is directly related to the maximum number of dealers.

$$Q_{max} \equiv Q(E=0) = \alpha N_{max}^{\beta} = \alpha^{1/(1-\beta)}\left(\frac{\pi}{w_0}\right)^{\beta/(1-\beta)}. \qquad (4.33)$$

Perhaps the most interesting thing about this expression is that sales increase more than linearly in the demand proportionality constant. That is, doubling demand by doubling α will more than double sales. The reason for this surprising result is simple. Increasing demand increases the number of sales which increases dealers' profits and attracts more dealers. Increasing the number of dealers dilutes enforcement pressure making the market more attractive to dealers. As still more dealers enter, sales increase still further.

The key to this feedback is that dealers' costs decrease as the market grows. This is an economy of scale, and economies of scale can give rise to downward sloping supply curves. Chapter 7 discusses some of the interesting implications of downward sloping supply curves for illicit drugs.

Equation 4.33 also shows that no matter what the demand parameter β is,

$$Q_{max} = \frac{w_0}{\pi} N_{max}. \qquad (4.34)$$

Thus the volume of sales in a market before a crackdown is proportional to the number of dealers. Furthermore, the proportionality constant is probably the same for all markets in a city.

The reservation wage w_0 is certainly the same for all markets in a city. The only question is whether the profitability per transaction π is as well. Generally the profitability per transaction, π, would be as well. It might not be if, for example, the market of interest is unusually violent. Then π might be smaller than it is in other markets. Equation 4.34 suggests that dealers in such a market conduct more transactions than do dealers in other markets. This is only reasonable; they must be compensated for the extra risk of violence or they would leave the market.

If π is constant throughout the city, Equation 4.34 suggests that before the crackdown, the volume of sales in all markets is roughly proportional to the number of dealers in that market. If one market has twice as many dealers as another, it probably also generates about twice as many sales.

That result is important because, while it is difficult to measure the number of dealers, it is next to impossible to measure directly the volume of sales.[56] Equation 4.34 gives an indirect way to measure sales.

Equation 4.34 also says that when the markets are all in equilibrium before a crackdown, dealers everywhere make about the same number of sales. This makes intuitive sense, and hence serves as an informal check that the model behaves as it should.

4.6.2 The Amount of Effort Needed to Collapse a Market

The quantity E_{max} is the effort needed to make

$$\underset{N>0}{\text{Max}}\{F(N)\} = 0, \tag{4.35}$$

i.e. to make $F(\overline{N}) = 0$. This is the value of E such that

$$c_1\left[\pi\,\alpha\left(\frac{\gamma\,E^{\gamma}}{\pi\,\alpha\,(1-\beta)}\right)^{\frac{\beta-1}{\gamma+\beta-1}} - E^{\gamma}\left(\frac{\gamma\,E^{\gamma}}{\pi\,\alpha\,(1-\beta)}\right)^{\frac{-\gamma}{\gamma+\beta-1}} - w_0\right] = 0.$$

[56] Kleiman (1988a, p.5) notes that measuring consumption citywide is quite difficult; determining the fraction of drugs coming from various markets within a city is that much harder.

$$(4.36)$$

The solution is

$$E_{max} = \left(\frac{1-\beta}{\gamma - 1+\beta}\right)^{1/\gamma}\left(\frac{\gamma - 1+\beta}{\gamma}\right)^{1/(1-\beta)} w_0{}^{1/\gamma}\ N_{max}. \tag{4.37}$$

This reduces to

$$= (1-\beta)\,\beta^{\beta/(1-\beta)}\, w_0\ N_{max} \tag{4.38}$$

for $\gamma = 1$.

Since w_0 is constant across all markets in a city, Equation 4.37 implies the amount of effort needed to collapse a market is proportional to the market's size. Combining Equations 4.34 and 4.37 gives

$$E_{max} = \left(\frac{1-\beta}{\gamma - 1+\beta}\right)^{1/\gamma}\left(\frac{\gamma - 1+\beta}{\gamma}\right)^{1/(1-\beta)} w_0{}^{(1-\gamma)/\gamma}\,\pi\ Q_{max} \tag{4.39}$$

so this result holds whether market size is measured in terms of the number of dealers or the volume of sales.

That the effort needed to collapse a market is proportional to its size is certainly plausible, but a priori other results would have been plausible too. Without the balloon model it would be hard to argue persuasively that the effort needed is not proportional to the square of the market's size or to the size of the market raised to some other power.

If the effort required is proportional to the size of the market it makes sense to speak of a critical ratio of enforcement pressure to market size as Kleiman[57] hypothesized would be the case. Kleiman notes that the Lynn crackdown, which he considers a success, involved about one officer for every 75 users. In contrast, the Lawrence crackdown, which seems not to have achieved lasting results, involved about one officer for every 150 users. It may be that the ratio of E_{max} to the size of the market for markets like those (assuming they are similar) is between 1/150 and 1/75 when measured in these units.

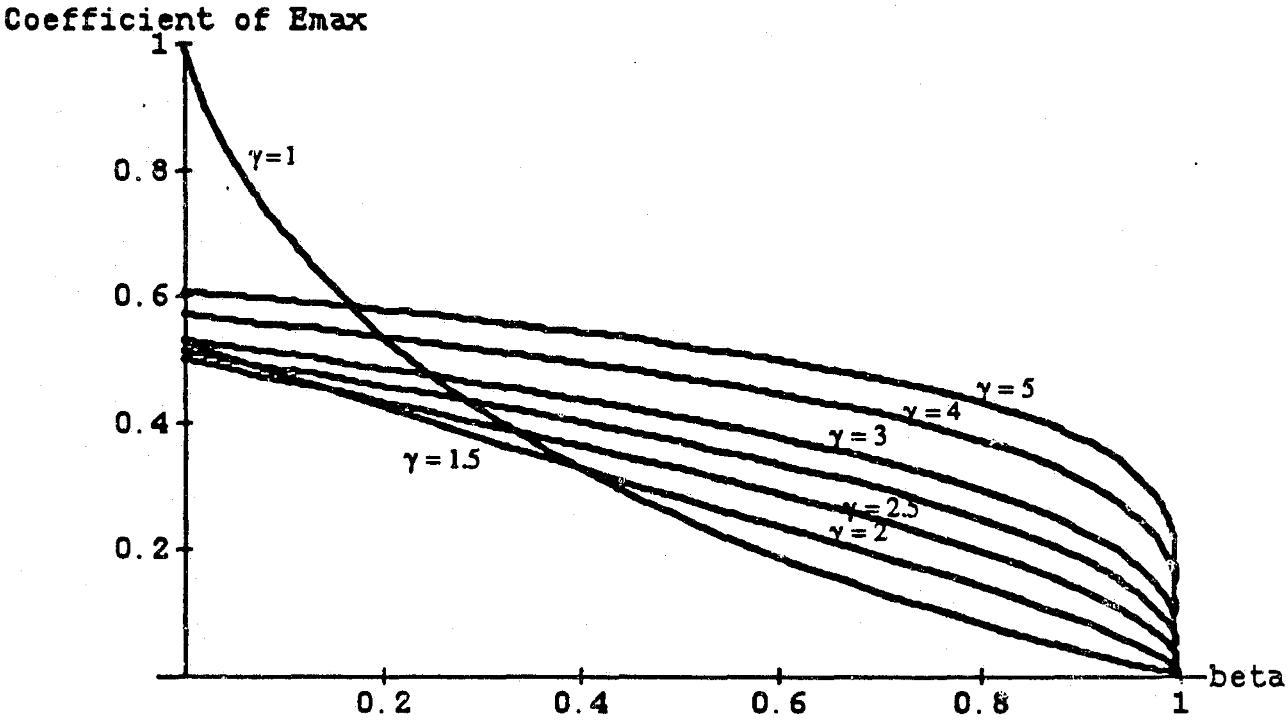
[57] Kleiman, 1988a, pp.25-26 and p.29.

The proportionality constant in the model is quite complicated. It depends on the demand parameter β, the risk aversion coefficient γ, the reservation wage w_0 and, in the case of Equation 4.39, on π as well. The reservation wage w_0 is the same for all markets in a city. Since dealers move between markets the risk aversion parameter γ probably is too, and, as discussed above, π is probably also constant throughout the city. Only β is likely to vary from market to market, and hence the only part of the coefficient that is likely to vary is

$$C(\beta,\gamma) \equiv \left(\frac{1-\beta}{\gamma - 1+\beta}\right)^{1/\gamma}\left(\frac{\gamma - 1+\beta}{\gamma}\right)^{1/(1-\beta)}. \tag{4.40}$$

Figure 4.9 plots this expression as a function of β for various values

Figure 4.9:
Coefficient of E_{max}, $C(\beta,\gamma)$

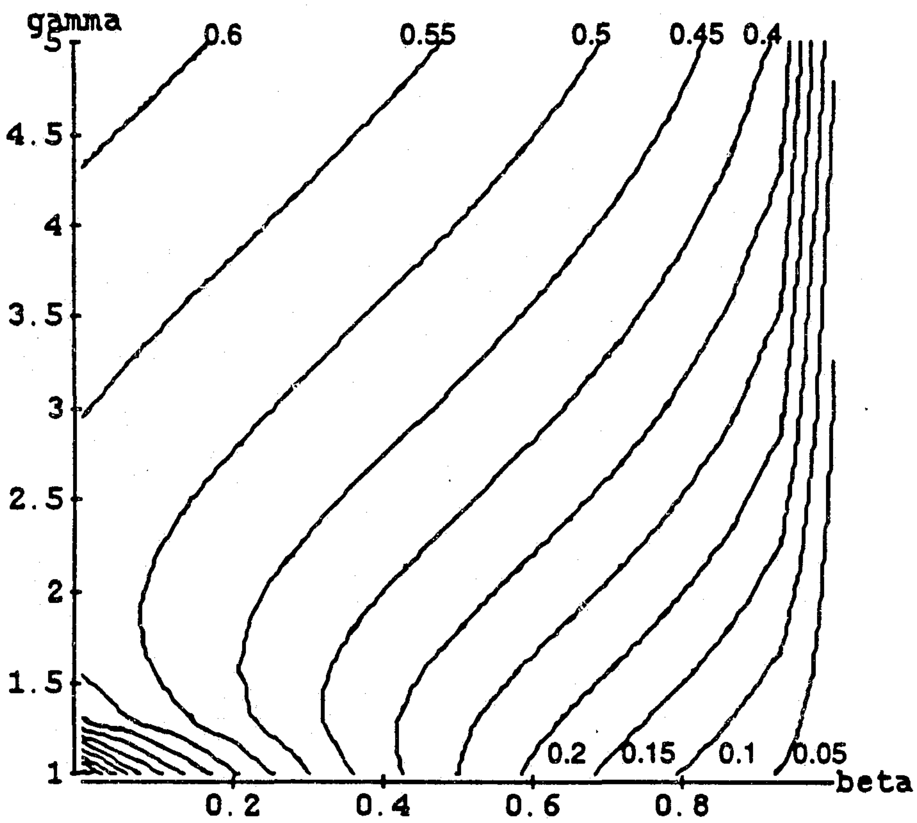


of γ. These plots show how the ratio of enforcement to market size needed to collapse the market might vary from market to market.

Figure 4.10 displays the same information in a different way; it plots the level curves of C(β,γ) on the β-γ plane.

The figures show that C(β,γ), and hence the ratio of E_{max} to the size of the market, is greatest for smaller values of β. When γ is not too large, the variation can be substantial. Consider how this information could be used. Suppose police are confronted with two markets of the same size, but one of the markets is a buyers' market (β small) and the other is a sellers' market (β closer to 1). Since C(β,γ) decreases in β, it would probably be easier, perhaps even much easier, to collapse the second market.

Figure 4.10:
Level Curves of C(β,γ)
(Labels of isoquants are approximate)



This suggests following the maxim "Go for the weak link." For markets of a given size, enforcement aimed at dealers will be more effective if availability of dealers is the limiting factor, i.e. if β is close to 1.

The result above is also just one of many in which the demand parameter β plays a decisive role. This raises the question of how β might be measured. Section 4.10 suggests one possibility, but it would be valuable to find other approaches.

4.6.3 The Minimum Viable Market Size

The value of N_{min} is just $N(E_{max})$, i.e. the value of N for which $F(N) = 0$ when $E = E_{max}$. This value is

$$N_{min} = \left(1 - \frac{1-\beta}{\gamma}\right)^{1/(1-\beta)} \left(\frac{\pi\,\alpha}{w_0}\right)^{1/(1-\beta)} \qquad (4.41)$$

$$= \left(1 - \frac{1-\beta}{\gamma}\right)^{1/(1-\beta)} N_{max}.$$

So N_{min} is proportional to N_{max}. That is, for given values of β and γ, no matter what the market's original size, it will have to be reduced by the same fraction before it collapses.

This is true whether size is measured in terms of the number of dealers or the volume of sales since

$$Q_{min} = \alpha\, N_{min}{}^{\beta} = \alpha\left(1 - \frac{1-\beta}{\gamma}\right)^{\beta/(1-\beta)} N_{max}{}^{\beta}$$

$$= \left(1 - \frac{1-\beta}{\gamma}\right)^{\beta/(1-\beta)} Q_{max}. \qquad (4.42)$$

So Q_{min} is proportional to Q_{max}.

The proportionality constants differ. In particular, defining

$$n_{min} \equiv \frac{N_{min}}{N_{max}} = \left(1 - \frac{1-\beta}{\gamma}\right)^{1/(1-\beta)} \text{ and} \qquad (4.43)$$

$$q_{min} \equiv \frac{Q_{min}}{Q_{max}} = \left(1 - \frac{1-\beta}{\gamma}\right)^{\beta/(1-\beta)}, \qquad (4.44)$$

one can see that

$$q_{min} = n_{min}^{\beta}. \qquad (4.45)$$

Since $n_{min} < 1$ and $\beta < 1$, this says that just before the market collapses the volume of sales will be reduced by a smaller fraction than the number of dealers will be.

A moment's reflection reveals that this is really a consequence of the assumption that sales increase less than linearly in the number of dealers. As discussed above, this assumption seems quite reasonable. So one is left with the disturbing conclusion that in general, local operations probably actually accomplish less than it appears they do, at least if the underlying objective is to reduce sales and the surrogate measure for sales is the number of active dealers. In particular this suggests that drug-use related property crime probably falls by a smaller fraction than the number of dealers does.

Taking derivatives shows that both n_{min} and q_{min} are increasing in γ, although they are increasing at a decreasing rate. This is reasonable. The more risk averse the dealers, the less enforcement can shrink the market before it collapses.

Both n_{min} and q_{min} depend on β in more complex ways, as Figures 4.11 and 4.12 show. First of all, n_{min} appears to be increasing in β while q_{min} appears to be a decreasing function of β. To understand why this is so, think about a buyer's market in which there is a surplus of dealers (β is small). Then even after a great deal of pressure has been applied, sales will not necessarily decrease substantially because there was originally a surplus of dealers. So q_{min} is large for small β. In contrast, the enforcement pressure can be expected to be relatively successful at driving away dealers because it was a buyer's market and hence was relatively unappealing to the dealers. So n_{min} is small for small β.

Equation 4.34 indicates the average number of sales a dealer makes per day when there is no (extra) enforcement ($N = N_{max}$ and $Q = Q_{max}$) is $\frac{w_0}{\pi}$. By Equations 4.41 and 4.42, just before enforcement bursts the market

$$\frac{Q}{N} = \frac{Q_{min}}{N_{min}} = \left(\frac{\gamma}{\gamma + \beta - 1}\right)\frac{w_0}{\pi} > \frac{w_0}{\pi}. \qquad (4.46)$$

Fig 4.11:
n_{min} as a Function of β for Various γ

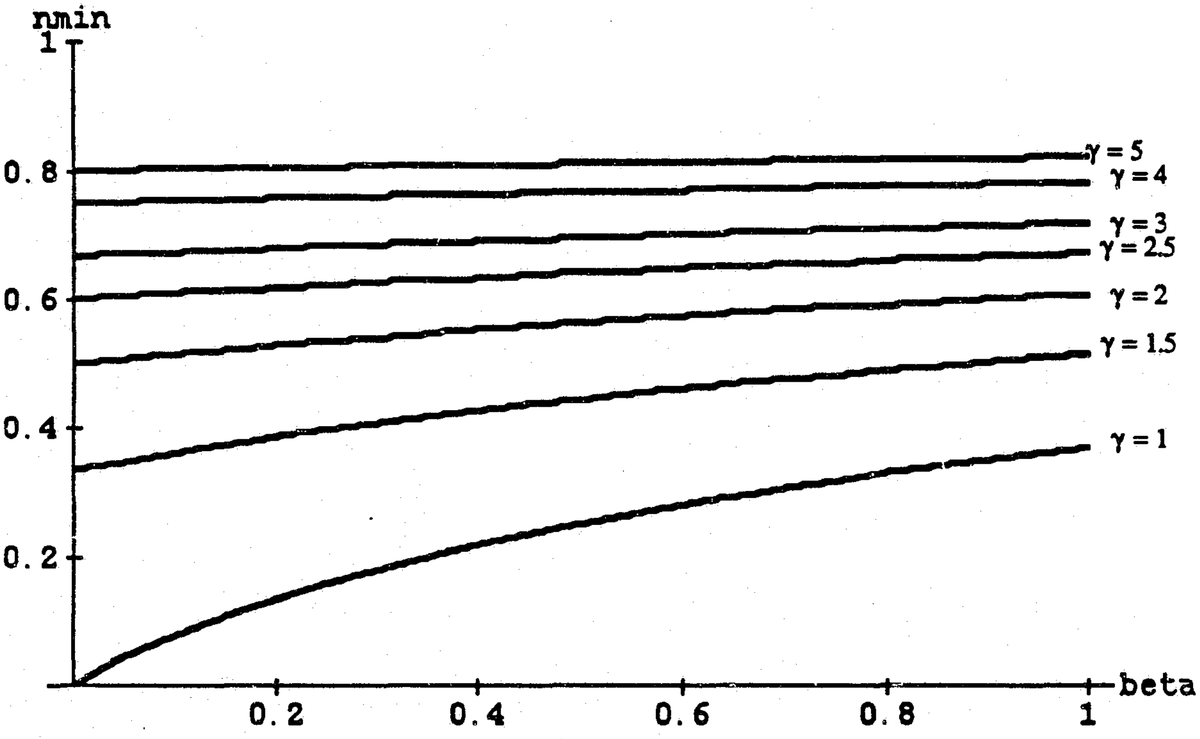
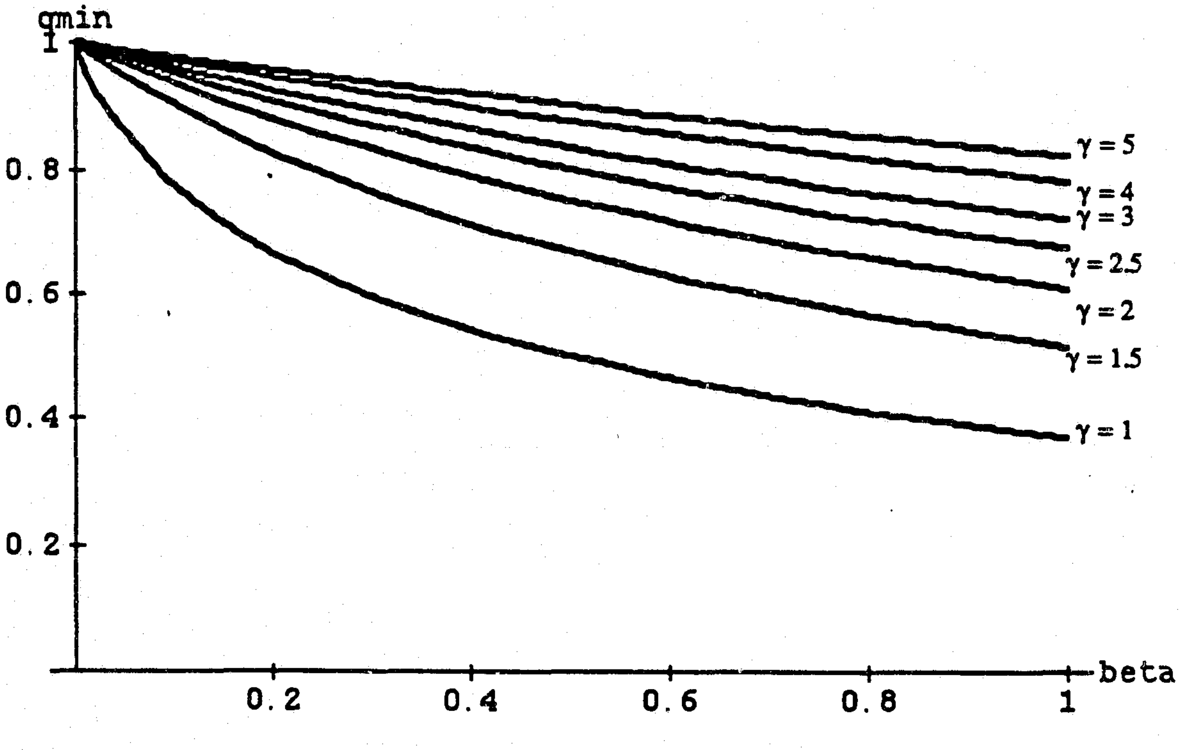


Fig 4.12:
q_{min} as a Function of β for Various γ



Thus in a market that is squeezed by local enforcement but not yet burst, one would expect to see fewer dealers, but these dealers would each be conducting more business and making more money than they were before the crackdown. This makes sense because dealers in that market must be compensated for the additional risk they incur. If dealers are risk averse and/or it is a sellers' market (β close to 1), the number of sales per dealer will be larger when enforcement pressure is applied, but not much larger. However, if dealers are almost risk neutral and the market was originally a buyers' market (β small), the relatively few dealers who remain active just before the market collapses make many more sales per day than they did originally.

This section has shown that E_{max} is proportional to N_{max} and Q_{max}; N_{min} is proportional to N_{max}; and Q_{min} is proportional to Q_{max}. Thus the size of the market does not affect how far one must push down to collapse the market; rather it is the nature of the demand and risk aversion, i.e. of β and γ.

For a market of a given size, the smaller β is, the harder one must work to collapse the market. Also, the smaller β is, the farther one must push down before the market collapses when size is measure in terms of the number of dealers; however, when market size is measured in terms of the volume of sales, one has to push down farther to make the market collapse when β is large.

It is difficult to say how the effort needed to collapse the market depends on the risk aversion parameter γ for two reasons. First, the coefficient $C(\beta,\gamma)$ is not monotonic in γ. Second, $E_{max} = C(\beta, \gamma)\, w_0^{\gamma}\, N_{max}$, and w_0 is not known quantitatively.

However, it is true that the more risk averse dealers are, the less the market can shrink before it collapses, no matter whether market size is measured in terms of the number of dealers or the volume of sales.

4.6.4 General Solution of E(N)

Although one cannot solve for N(E) for arbitrary β and γ, one can solve for the inverse function E(N). In equilibrium

$$\pi \alpha N^{\beta-1} - \left(\frac{E}{N}\right)^{\gamma} - w_0 = 0, \qquad (4.47)$$

so

$$E = \left(\pi \alpha N^{\beta-1} - w_0\right)^{1/\gamma} N \tag{4.48}$$

for $N_{min} \leq N \leq N_{max}$. Using the definition $N \equiv n N_{max}$, this implies

$$E = \left(n^{\gamma+\beta-1} - n^{\gamma}\right)^{1/\gamma} w_0{}^{1/\gamma} N_{max}. \tag{4.49}$$

Then Equation 4.37 for E_{max} and the definition $E \equiv e_N E_{max}$, imply

$$e_N = \left(\frac{\gamma + \beta - 1}{1 - \beta}\right)^{1/\gamma} \left(\frac{\gamma}{\gamma + \beta - 1}\right)^{1/(1-\beta)} \left(n^{\gamma+\beta-1} - n^{\gamma}\right)^{1/\gamma}. \tag{4.50}$$

Figure 4.13a-e piots e(n) for $\beta = 0.1, 0.3, 0.5, 0.7$, and 0.9 for $\gamma = 1.0$, 1.5, 2.0, 2.5, and 3.0. In all five graphs, the smaller the value of β, the faster the curve increases for small n.

Figure 4.13a:
$e_N(n)$ for $\gamma = 1.0$ and $\beta = 0.1, 0.3, 0.5, 0.7$, and 0.9

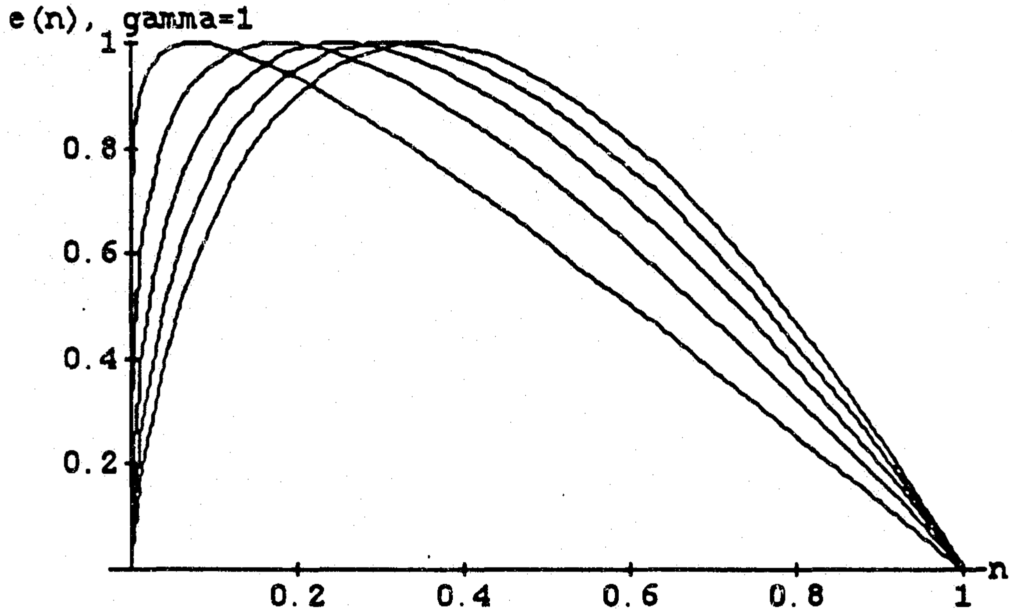


Figure 4.13b:
$e_N(n)$ for $\gamma = 1.5$ and $\beta = 0.1, 0.3, 0.5, 0.7,$ and 0.9

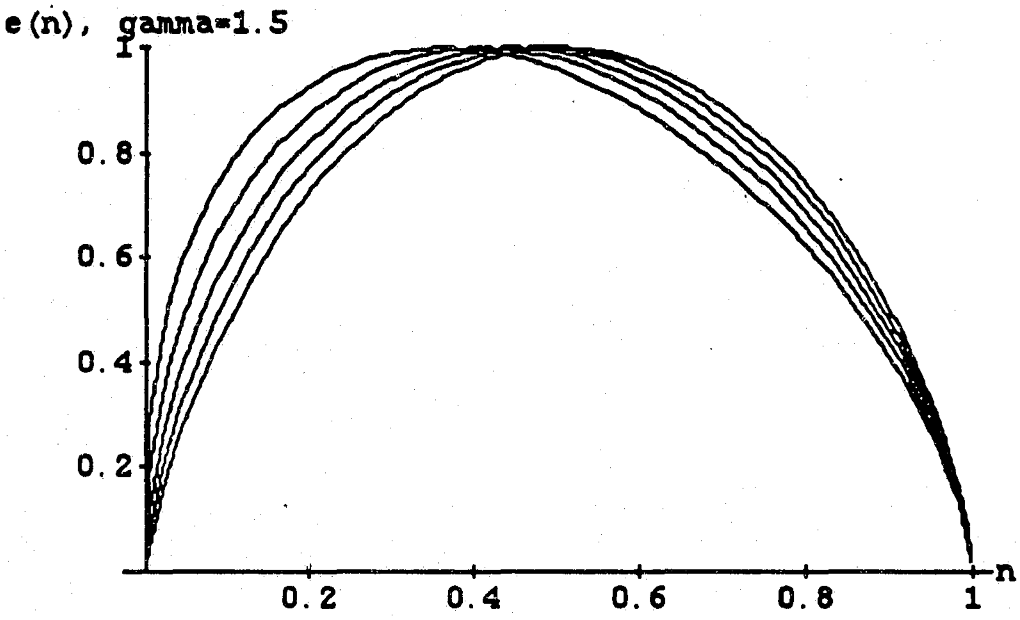


Figure 4.13c:
$e_N(n)$ for $\gamma = 2.0$ and $\beta = 0.1, 0.3, 0.5, 0.7,$ and 0.9

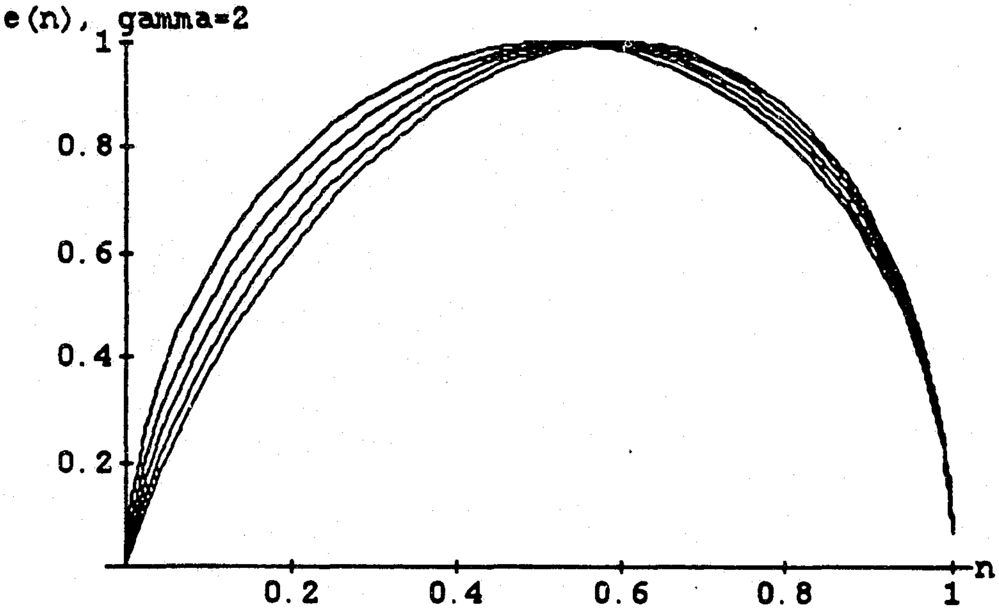


Figure 4.13d:
$e_N(n)$ for $\gamma = 2.5$ and $\beta = 0.1, 0.3, 0.5, 0.7,$ and 0.9

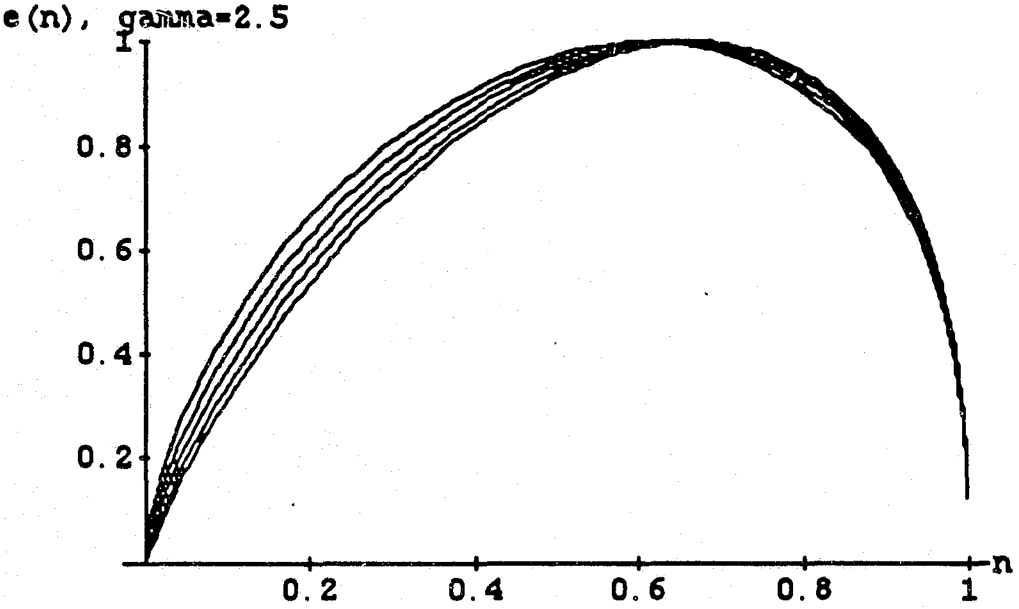
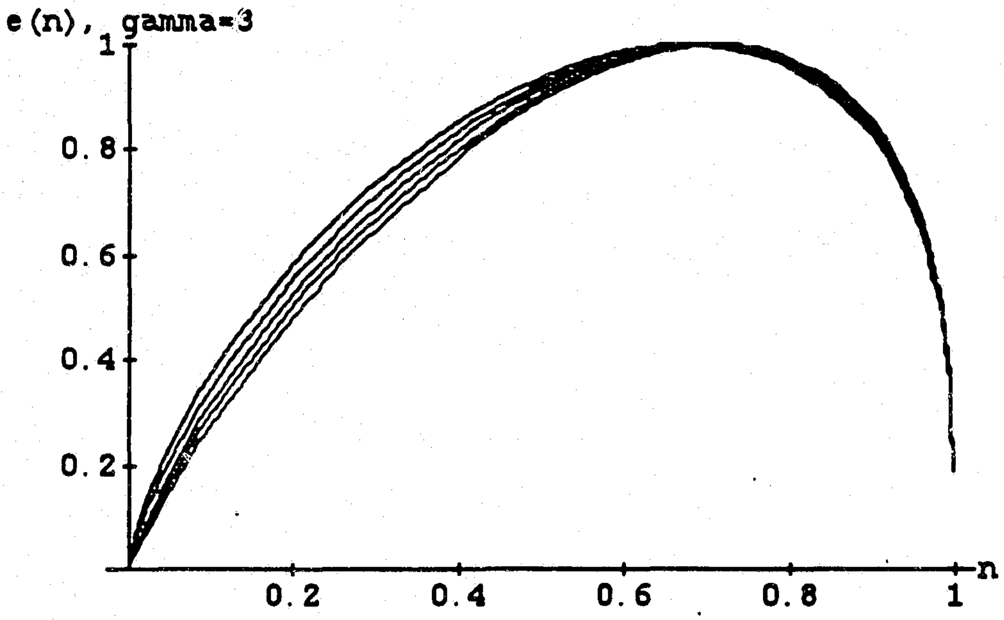


Figure 4.13e:
$e_N(n)$ for $\gamma = 3.0$ and $\beta = 0.1, 0.3, 0.5, 0.7,$ and 0.9



Each curve has two parts. The part to the left of the peak does not represent stable equilibria; its interpretation will be given in Section 4.8. The part to the right of the peak is the locus of stable equilibria; it shows which market-size/enforcement pairs constitute long run equilibria. Inverting the axes gives n as a function of e. Note that the graphs are not directly comparable because the normalizations for e are different for different values of β and γ.

The plots are still useful, however. They show, for instance, that as the risk aversion parameter γ increases, the demand parameter β becomes less important, as was suggested earlier by the plots of n_{min} and q_{min}. They also show that the positive feedback effect is strongest for large β, and that when γ is large there will be little reduction in the number of dealers until enforcement exceeds about half that which is required to collapse the market.

Consider next Equation 4.49. For two markets of the same size in a one city, $w_0^{1/\gamma}N_{max}$ is a constant. Hence the effort required to reduce the size of the market (measured in terms of the number of dealers) for given values of β and γ is determined by $\left(n^{\gamma+\beta-1} - n^{\gamma}\right)^{1/\gamma}$. Denote this expression by $e(n,\beta,\gamma)$. Taking derivatives shows that

$$\frac{de(n,\beta,\gamma)}{d\beta} < 0, \tag{4.51a}$$

$$\mathrm{Sgn}\left(\frac{d^2e(n,\beta,\gamma)}{d\beta^2}\right) = \mathrm{Sgn}\left(n^{\beta-1} - \gamma\right), \tag{4.51b}$$

$$\mathrm{Sgn}\left(\frac{de(n,\beta,\gamma)}{dn}\right) < 0, \tag{4.51c}$$

$$\frac{d^2e(n,\beta,\gamma)}{dn^2} < 0, \tag{4.51d}$$

$$\mathrm{Sgn}\left(\frac{de(n,\beta,\gamma)}{d\gamma}\right) = \mathrm{Sgn}\left(n - \left(\frac{1}{2}\right)^{1/(1-\beta)}\right), \text{ and} \tag{4.51e}$$

$$\frac{d^2e(n,\beta,\gamma)}{d\gamma^2} < 0 \quad \text{iff} \quad \left(\frac{1}{2}\right)^{1/(1-\beta)} < n < \left(\frac{1}{1+e^{-2\gamma}}\right)^{1/(1-\beta)}. \tag{4.51f}$$

Some of these derivatives can be translated into simple English statements. For instance, Equation 4.51a confirms that when β is small, one needs to exert a larger fraction of the effort needed to collapse the market to achieve a given reduction in the size of the market. This means that when β is small, a crackdown could be close to collapsing a market before much progress is apparent. In contrast, in a market with $\beta \approx 1$, if little progress has been made after a significant fraction of the available resources have been applied, it is less likely that the market will collapse even if all available enforcement resources are focused on that market. In that case it might be wise to choose a smaller target.

This distinction is important because several researchers have advocated taking a "try it and see" approach.[58] If β is large this seems sensible. If β is small it may still be a good idea. If the trial program collapses the market, one would have a definitive answer about the effectiveness of the program. But if β is small and the crackdown does not make a substantial dent in the market, one cannot safely say that a modest increment in effort would not be enough to collapse the market.

Equation 4.51c simply says that the more enforcement pressure that is applied, the smaller the market will be. Equation 4.51d, however, is much more significant. It generalizes Equation 4.24 demonstrating that there is positive feedback for all values of γ and all $\beta \in (0,1)$.

The derivatives with respect to γ are harder to interpret, and they are less important because γ would generally be constant throughout the city.

4.7 Some Explicit Solutions for N(E)

The previous section derived results that held for all $\beta \in (0,1)$ and all $\gamma \geq 1$. The results obtained included most of what one would want to know except for an explicit expression for N(E). There is no simple closed form solution for N(E) for all γ and β, but there are solutions if $\gamma = 1 - \beta$, $4(1-\beta)/3$, $3(1-\beta)/2$, $2(1-\beta)$, $3(1-\beta)$, and $4(1-\beta)$ and for all γ when $\beta = 1$.

Subsection 4.5.1 gives the results for $\beta = 1$ and $\gamma = 1$. The generalization to $\beta = 1$ and $\gamma > 1$ is trivial and uninformative.

[58] Kleiman (1988a) and Barnett (1988).

Finding N(E) when $\gamma = 4(1-\beta)/3$ and $\gamma = 4(1-\beta)$ requires solving a quartic expression. This is possible of course, but the algebra required is daunting.

The results for $\gamma = 1 - \beta$ and $\gamma = 2(1-\beta)$ are straightforward generalizations of the results in Subsections 4.5.2 and 4.5.3, respectively. The truly new material presented here are the solutions for $\gamma = 3(1-\beta)/2$ and $\gamma = 3(1-\beta)$. Subsection 4.7.5 plots N(E) for various values of β and γ; the plots graphically illustrate the conclusion obtained above that for a given size market, less effort is required to collapse the market if β is large.

4.7.1 Solution for $\gamma = 1 - \beta$

For $\beta \in [0,1)$ and $\gamma = 1 - \beta$, setting $\frac{dN}{dt}$ equal to zero implies

$$w_0 N^{\gamma} - \pi \alpha N^{\gamma+\beta-1} + E^{\gamma} = 0. \tag{4.52}$$

The steady state solution has

$$N = \left(\frac{\pi \alpha - E^{\gamma}}{w_0}\right)^{1/\gamma}. \tag{4.53}$$

Using Equations 4.32, 4.33, and 4.37 and the definitions of n, q, and e_N, Equation 4.53 can be manipulated to obtain

$$n = 1 - e_N{}^{\gamma} \text{ and} \tag{4.54a}$$

$$q = (1 - e_N{}^{\gamma})^{1-\gamma}. \tag{4.54b}$$

4.7.2 Solution for $\gamma = 2(1-\beta)$

The solution for $\beta \in (0,1)$ and $\gamma = 2(1 - \beta)$ is a straightforward generalization of the solution for $\beta = {}^1/_2$ and $\gamma = 1$ discussed in Subsection 4.5.3. Setting $\frac{dN}{dt}$ equal to zero implies

$$w_0 N^{\gamma} - \pi \alpha N^{\gamma/2} + E^{\gamma} = 0. \tag{4.55}$$

The steady state solution is

$$N = \left(\frac{\pi \alpha}{2 w_0} + \frac{1}{2}\sqrt{\left(\frac{\pi \alpha}{w_0}\right)^2 - \frac{4E^{\gamma}}{w_0}}\right)^2. \tag{4.56}$$

Again using Equations 4.32, 4.33, and 4.37 and the definitions of n, q, and e_N, this can be manipulated to obtain

$$n = \left[\frac{1}{2}\left(1 + \sqrt{1 - e_N{}^{\gamma}}\right)\right]^{2/\gamma} \text{ and } \tag{4.57a}$$

$$q = \left[\frac{1}{2}\left(1 + \sqrt{1 - e_N{}^{\gamma}}\right)\right]^{\frac{2}{\gamma} - 1}. \tag{4.57b}$$

With explicit solutions for N(E), one can also write an expression for the number of additional dealers the police must remove to make the market collapse at any given level of enforcement E. This expression is just $N^* - \tilde{N}$. For $\gamma = 2(1 - \beta)$ it is

$$\left(\frac{\pi\,\alpha}{2\,w_0} + \frac{1}{2}\sqrt{\left(\frac{\pi\,\alpha}{2\,w_0}\right)^2 - \frac{4E^{\gamma}}{w_0}}\right)^{2/\gamma} - \left(\frac{\pi\,\alpha}{2\,w_0} - \frac{1}{2}\sqrt{\left(\frac{\pi\,\alpha}{2\,w_0}\right)^2 - \frac{4E^{\gamma}}{w_0}}\right)^{2/\gamma}$$

$$= N_{max}\left(\left[\frac{1}{2}\left(1 + \sqrt{1 - e_N{}^{\gamma}}\right)\right]^{2/\gamma} - \left[\frac{1}{2}\left(1 - \sqrt{1 - e_N{}^{\gamma}}\right)\right]^{2/\gamma}\right). \tag{4.58}$$

For $\gamma = 1$ or 2 this reduces to

$$= N_{max}\sqrt{1 - e_N{}^{\gamma}}. \tag{4.59}$$

4.7.3 Solution for $\gamma = 3(1-\beta)$

For $\beta \in (0,1)$ and $\gamma = 3(1 - \beta)$, setting $\frac{dN}{dt}$ equal to zero implies

$$w_0\,N^{\gamma} - \pi\,\alpha\,N^{\gamma+\beta-1} + E^{\gamma} = 0. \tag{4.60}$$

The steady state solution can be obtained by solving a cubic equation. It is

$$N = \left(\frac{\pi\,\alpha}{3\,w_0}\left[1 + 2\,\mathrm{Cos}\left[\frac{1}{3}\,\mathrm{Cos}^{-1}\left(1 - 2e_N{}^{\gamma}\right)\right]\right]\right)^{3/\gamma}. \tag{4.61}$$

This implies

$$n = \left(\frac{1}{3} + \frac{2}{3}\,\mathrm{Cos}\left[\frac{1}{3}\,\mathrm{Cos}^{-1}\left(1 - 2e_N{}^{\gamma}\right)\right]\right)^{3/\gamma}, \text{ and} \tag{4.62a}$$

$$q = \left(\frac{1}{3} + \frac{2}{3}\mathrm{Cos}\left[\frac{1}{3}\mathrm{Cos}^{-1}(1 - 2e_N{}^{\gamma})\right]\right)^{\frac{3}{\gamma}-1}. \qquad (4.62b)$$

4.7.4 Solution for $\gamma = 3(1-\beta)/2$

Finally, for $\beta \in (0,1)$ and $\gamma = 3(1 - \beta)/2$ setting $\frac{dN}{dt}$ equal to zero implies

$$w_0 N^{\gamma} - \pi\alpha N^{\gamma+\beta-1} + E^{\gamma} = 0 \qquad (4.63)$$

Again the steady state solution is obtained by solving a cubic equation. The result is

$$N = \left(2\sqrt{\frac{\pi\alpha}{3 w_0}}\ \mathrm{Cos}\left[\frac{1}{3}\mathrm{Cos}^{-1}(-e_N{}^{\gamma})\right]\right)^{3/\gamma}, \qquad (4.64)$$

so

$$n = \left(\frac{2}{\sqrt{3}}\ \mathrm{Cos}\left[\frac{1}{3}\mathrm{Cos}^{-1}(-e_N{}^{\gamma})\right]\right)^{3/\gamma} \text{ and} \qquad (4.65a)$$

$$q = \left(\frac{2}{\sqrt{3}}\ \mathrm{Cos}\left[\frac{1}{3}\mathrm{Cos}^{-1}(-e_N{}^{\gamma})\right]\right)^{\frac{3}{\gamma}-2}. \qquad (4.65b)$$

4.7.5 Graphical Comparison of N(E) for Various γ and β

Comparing these solutions graphically illustrates the crucial role played by the demand parameter β. This cannot be done directly by plotting Equations 4.54a, 4.57a, 4.62a, and 4.65a, however, because each one is normalized differently. In each equation e is defined as E/E_{max} for the value of E_{max} given by Equation 4.37 for the particular values of β and γ. This is easy to correct though by defining a universal normalization

$$\tilde{e_N} = \frac{E}{w_0{}^{1/\gamma}\ N_{max}}$$

$$= \left(\frac{1-\beta}{\gamma+\beta-1}\right)^{1/\gamma}\left(\frac{\gamma+\beta-1}{\gamma}\right)^{1/(1-\beta)} e_N. \qquad (4.66)$$

Figure 4.14a plots N as a function of E for $\gamma = 1$ and $\beta = 0$, 1/3, 1/2, and 2/3. When E = 0 the number of dealers is the same for all β, but as E increases, N decreases faster for larger β. Figure 4.14b is the corresponding graph depicting Q as a function of E.

Figure 4.14a:
$N(\tilde{e}_N)$ for $\gamma = 1$ and $\beta = 0$, 1/3, 1/2 and 2/3

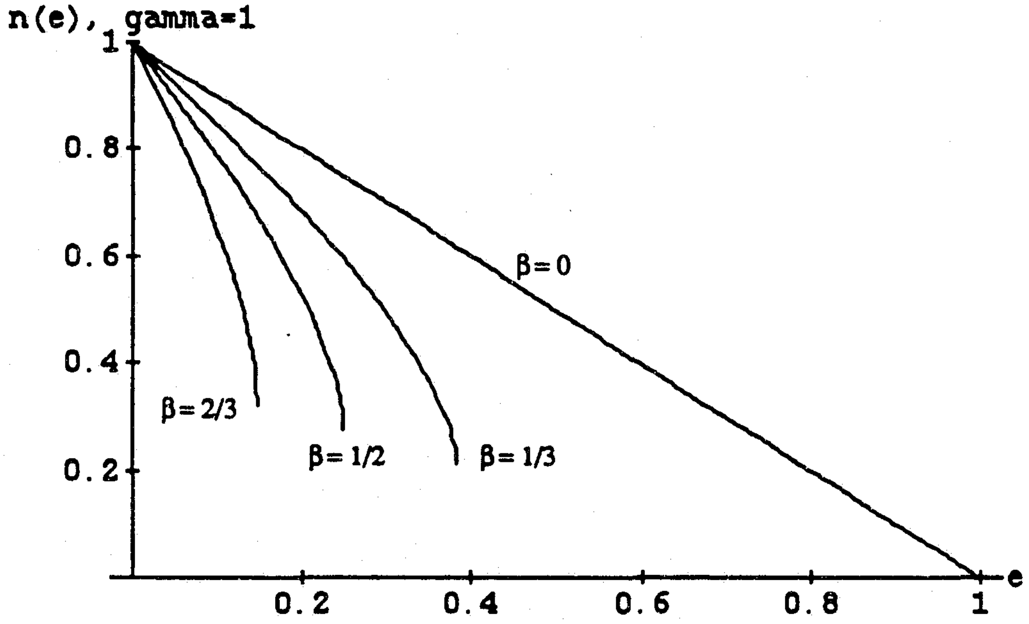
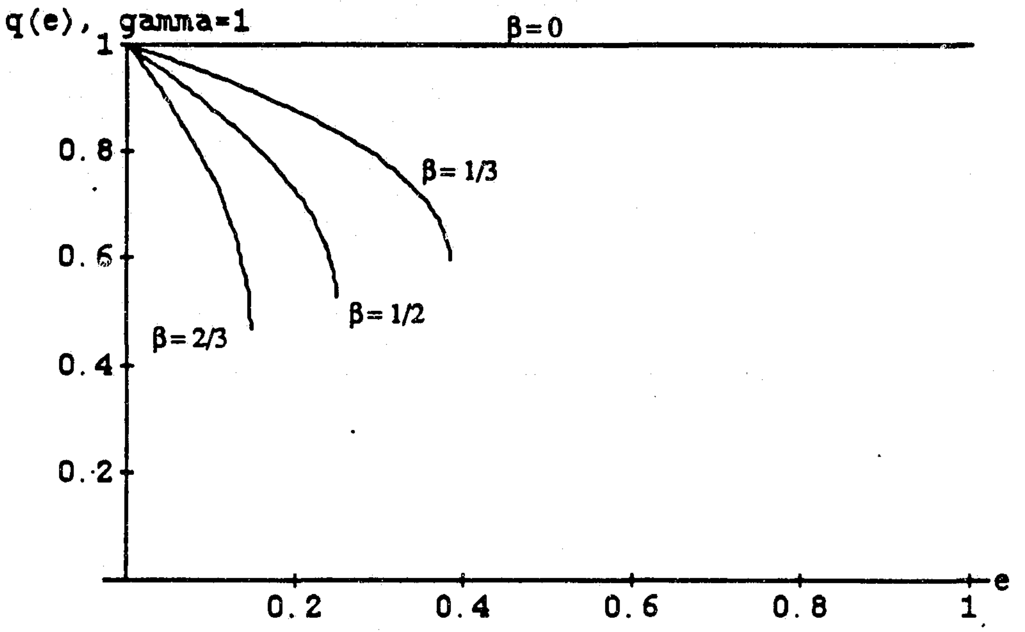


Figure 4.14b:
$Q(\tilde{e}_N)$ for $\gamma = 1$ and $\beta = 0$, 1/3, 1/2 and 2/3



Figures 4.15a,b are the corresponding plots for β = 0, 1/4, and 1/2 and γ = 3/2.

Figure 4.15a:
$N(\widetilde{e}_N)$ for $\gamma = 1.5$ and $\beta = 0, 1/4, 1/2$

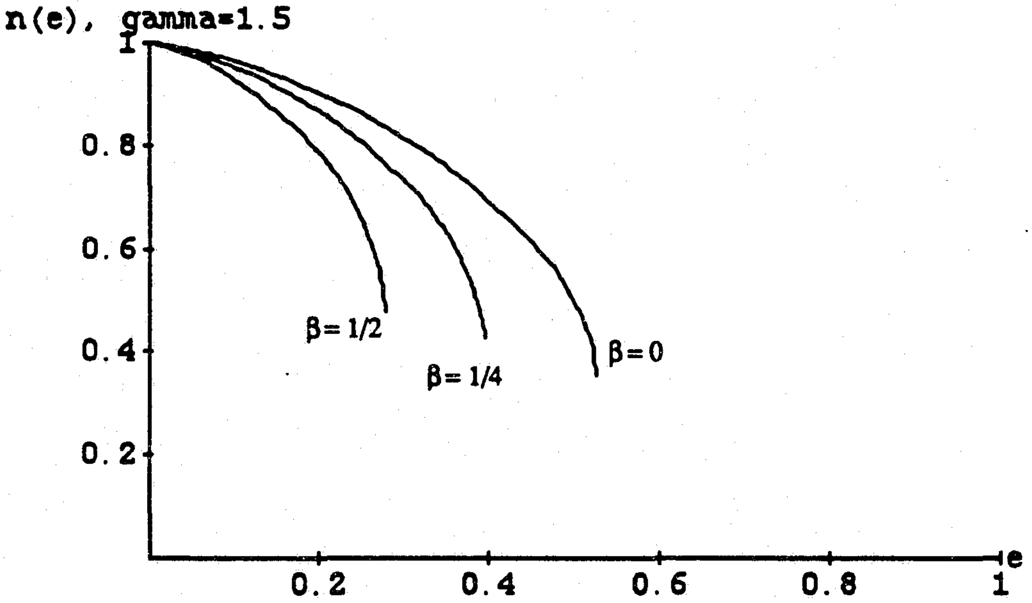
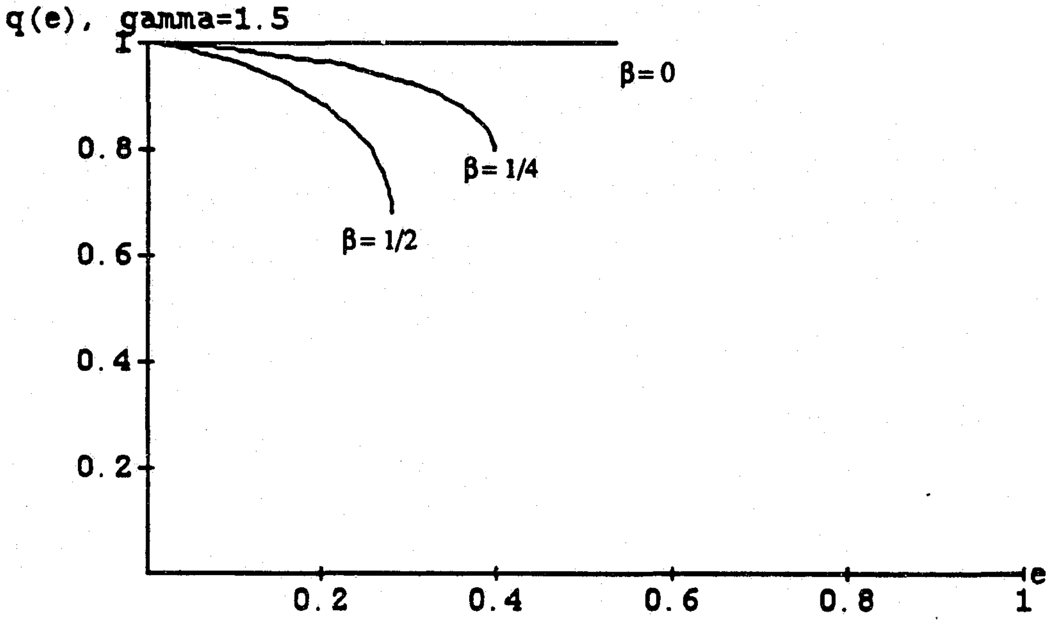


Figure 4.15b:
$Q(\widetilde{e}_N)$ for $\gamma = 1.5$ and $\beta = 0, 1/4, 1/2$



Figures 4.16a,b are the corresponding plots for $\beta = 0$ and 1/3 and for $\gamma = 2$.

Figure 4.16a:
$N(\tilde{e}_N)$ for $\gamma = 2.0$ and $\beta = 0, 1/3$

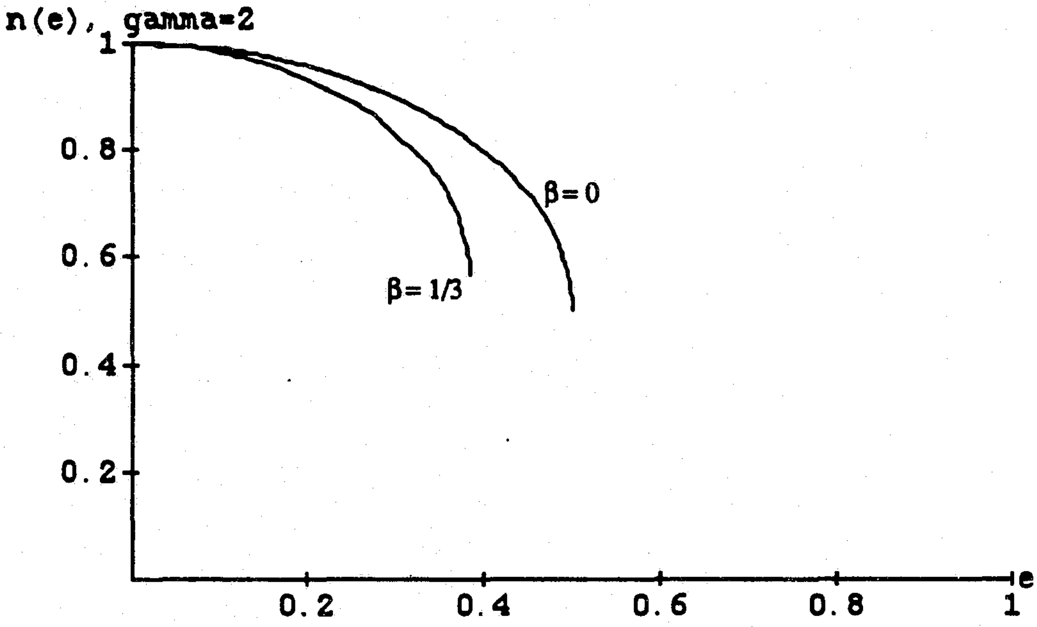
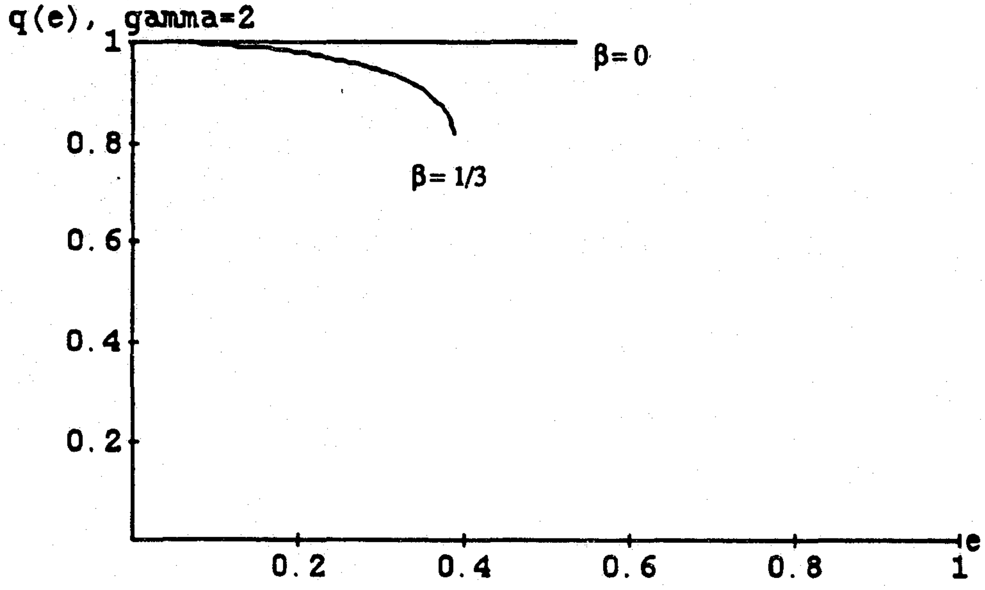


Figure 4.16b:
$Q(\tilde{e}_N)$ for $\gamma = 2.0$ and $\beta = 0, 1/3$



Two things are apparent from these plots. First, the positive feedback effect is substantial. Second, given two markets of the same size, it takes less effort to collapse the one with a larger demand parameter β, sometimes much less.

4.8 Will the Market Spring Back After a Crackdown?

A key question about crackdowns is, "Do they produce lasting results?" To be more specific, after a "successful" crackdown has eliminated dealing in a market, will the market spring back when the extra pressure applied during the crackdown is removed?

The answer described in Subsection 4.5.1 holds for all β and γ. Even a small amount of pressure can keep individuals from beginning to deal again because they would attract the full burden of the enforcement. On the other hand, if a group of dealers arrives at one time, they may be able to "jump start" the market. A key question then is: how much enforcement pressure must be maintained to prevent a given number of dealers from "jump starting" a market?

To see this let $\hat{N}$ be the number of dealers attempting to revive the market. If $\hat{N} \geq N_{min}$ then one needs to have enforcement E = E_{max}. If $\hat{N} < N_{min}$ less pressure is needed. The minimum amount needed is that level of enforcement for which $\frac{dN}{dt}$ is initially negative. That level is sufficient because as N decreases, $\frac{dN}{dt}$ becomes more negative.

So E must be large enough that

$$\frac{\pi \alpha \hat{N}^{\beta}}{\hat{N}} - \left(\frac{E}{\hat{N}}\right)^{\gamma} - w_0 < 0, \qquad (4.67)$$

i.e. that

$$E > \left[\pi \alpha \hat{N}^{\beta-1} - w_0\right]^{1/\gamma} \hat{N} \qquad (4.68)$$

$$= \left[\hat{n}^{\gamma+\beta-1} - \hat{n}^{\gamma}\right]^{1/\gamma} w_0^{1/\gamma} N_{max},$$

where $\hat{n} \equiv \frac{\hat{N}}{N_{max}}$. Using the definition $e_N \equiv \frac{E}{E_{max}}$, this implies that one needs

$$e_N > \left(\frac{\gamma + \beta - 1}{1 - \beta}\right)^{1/\gamma} \left(\frac{\gamma}{\gamma + \beta - 1}\right)^{1/(1-\beta)} \left[\hat{n}^{\gamma+\beta-1} - \hat{n}^{\gamma}\right]^{1/\gamma} \qquad (4.69)$$

$$= \frac{\hat{n}^{\beta} - \hat{n}}{(1 - \beta)\, \beta^{\beta/(1-\beta)}}$$

for $\gamma = 1$.

For any given market β and γ are constant, so the effort needed as a function of $\hat{n}$ is proportional to

$$\left[\hat{n}^{\gamma+\beta-1} - \hat{n}^{\gamma}\right]^{1/\gamma} = \left[\hat{n}^{\beta-1} - 1\right]^{1/\gamma} \hat{n}. \qquad (4.70)$$

What is most striking about this expression is how quickly the enforcement pressure needed increases as a function of $\hat{n}$. This has important implications. First, as discussed in Subsection 4.5.1, it helps explain why coordinating mechanisms such as gangs can catalyze the creation of drug markets. Note that gangs can play a decisive role in the creation of markets even if they do not ultimately control a large fraction of the sales.

To see this, suppose $\hat{n} = 0.2$ gang members jump start a market. Then as many as four times that many individual entrepreneurs will join the nascent market. So when someone observes a mature market they might incorrectly assume gangs played a minor role in creating that market because they are only making 20% of the sales.

This discrepancy is further compounded if the gang continues to colonize new markets. For a period after the gang first colonizes a market, its members earn wages exceeding the reservation wage; that is why other dealers enter. But once the market matures, all dealers make w_0. Actually gang members would probably continue to do better than non-members if membership offers some protection against dealer-dealer violence. Even if this is the case, it would still be true that because of competition in a mature market, the gang members would make less than they did when the market was growing.

This gives gangs an incentive to keep colonizing new markets. Hence, one gang that currently accounts for a relatively small fraction of the dealing in one fairly mature market may actually have been responsible for the creation of that market and others like it.

The second major implication of Equation 4.70 is that it is hard for police to carry out a successful crackdown campaign alone. Even if they manage to clean up several markets, they might have to allocate so much of the effort to maintenance, that they can no longer muster the strength needed to collapse the remaining markets.

Think of the situation in Hartford, where there are 22 markets. Assume for the moment, as is almost certainly not the case, that all 22 markets are the same size and have the same demand parameter β. Suppose a maintenance force equivalent to one-quarter of the effort needed to collapse a market is left in each collapsed market. Then in order to clean up all the markets the police would actually need six times the resources required to clean up one market.

Actually this little calculation oversteps the bounds of the model because the model applies only to crackdowns in one of many markets in a city. But intuition says that larger and larger maintenance forces would be necessary as the number of markets is reduced, so it may even understate the case.

Of course not all markets are the same size. Most of the results in this chapter suggest hitting the smallest and "easiest" markets first. Because of positive feedback, one accomplishes more by collapsing a small market than by denting a large one. However, Equation 4.70 suggests that if the long term goal is to clean up all the markets, one might not want to save the largest market for last.

To illustrate this, suppose there are two markets, one with $E_{max} = 2$ and one with $E_{max} = 1$. If the police attack the smaller market first, they will need one unit of strength to collapse the first market and a total of 2 1/4 units while collapsing the second (1/4 to maintain the first market plus 2 to collapse the second). If, on the other hand, they attack the larger market first, they will need 2 units of strength to collapse it and 1 1/2 units while collapsing the smaller one (1/2 to maintain the larger market plus 1 to collapse the smaller). Hence, the peak level of enforcement required is greatest if the largest market is saved for last.

If the police can not carry out a crackdown campaign alone, they need to enlist assistance from the community.[59] If community cohesiveness is restored, and people promptly report any resurgence of dealing in a collapsed market, then the police will not need to

[59] Moore and Kleiman (1990) discuss the need for police-community cooperation in confronting the drug problem.

maintain such a large presence. This point is one of the fundamental tenets of community policing literature.[60] This section backs up that insight with a quantitative argument.

Of course if the neighbors report some dealing the police should respond quickly. The model suggests that it is far easier to stamp out a small but growing market than it is to clean up a mature market.

Another implication is that police need to do crackdowns "nicely". In particular, they need to avoid aggravating racial tensions and alienation from authorities. If the police cannot complete a crackdown without damaging police-community relations, it is probably not worth undertaking the crackdown because the dealing could very well spring back.

4.9 Heterogeneity of Dealers

One of the model's stronger assumptions is that dealers are all identical. This section relaxes one aspect of that assumption by considering what would change if not all dealers had the same reservation wage.

Dealers might well prefer certain markets over others. For example, they might prefer to deal near their home because they know the terrain better, which might help them escape police pursuit; because being on their own turf protects them from dealer-dealer violence; or simply because it makes returning to their stash less inconvenient. Other factors that might be relevant are the ethnic composition of the neighborhood and the gang that claims the street as its territory. All of these things might lead some dealers to like dealing in a particular market more or less than other dealers do.

As a result, the reservation wage of different dealers in one market might be different. Consider, as an extreme example, a dealer who incurred the enmity of a powerful and violent person. That dealer might be highly dependent on the security afforded by remaining close to home and surrounded by friends, and so might have a lower reservation wage than others. On the other hand, a dealer who relies primarily on sales arranged through a beeper instead of selling to whatever customer happens to drive up might be able to switch markets at minimal cost and hence have a higher than average reservation wage.

[60] For an introduction to community policing see Kelling and Stewart (1989); Moore, Trojanowicz, and Kelling (1988); and Sparrow (1988).

If the reservation wages of the N dealers in the market are not all the same, then w_0 should be interpreted as the highest reservation wage among dealers in that market. It is probably fair to assume that some fraction of the dealers are mobile; they are equally happy dealing in any market. Their reservation wage determines w_0. The remainder of the dealers have varying intensities of preference for their current market. Those with only a mild preference have a reservation wage only slightly less than w_0. Those with a strong preference have a much lower reservation wage. Members of the last group are the ones least likely to be displaced by a crackdown; they are the most likely to respond to having their market shut down by abandoning dealing altogether.

When w_0 is interpreted as the highest reservation wage among dealers in the market, the basic principle of the balloon model still holds. If the utility derived by dealers in the market is less than w_0, then dealers will exit. If it exceeds w_0, dealers will enter.

Now imagine what happens when the police crack down. Let $\tilde{w}_0$ stand for the original reservation wage. As the enforcement per dealer rises, the wage falls below $\tilde{w}_0$. The first to exit are dealers whose reservation wage was $\tilde{w}_0$. If the enforcement pressure is not too severe and there are enough dealers with reservation wage $\tilde{w}_0$, then after some of them have left the market reaches the equilibrium that has been described throughout this chapter, and the reservation wage is still $\tilde{w}_0$.

If, on the other hand, the number of dealers that would have to leave to restore equilibrium ($N_{max} - N^*$) is greater than the number of dealers with a reservation wage $\tilde{w}_0$, then the reservation wage w_0 decreases. As the reservation wage decreases, the stable equilibrium size of the market increases. That is, the market equilibrium will have more dealers than would have been the case if all dealers had had a reservation wage equal to $\tilde{w}_0$.

Thus heterogeneity in reservation wages undercuts the positive feedback effect. With uniform reservation wages, the greater the effort level, the easier it is to push a given number of additional dealers out. If some dealers have lower reservations wages, however, then as the crackdown progresses it may become more and more difficult to dislodge additional dealers.

Hence the composition of the dealers in a market will affect the outcome of a crackdown. To further illustrate this point, suppose there are just two kinds of dealers. Type 0 dealers are limited to their own market, and hence have a low reservation wage. Type 1 dealers are mobile; they can operate in any market and hence have a higher reservation wage.

Suppose the police have decided to crack down on one of three markets. The first is occupied by mobile, Type 1 dealers; the second exclusively by immobile Type 0 dealers; the third has a mixture of dealers.

If the police crack down on the first, they have a good chance of collapsing it because the dealers' reservation wage is high. However, even if they collapse the market the total number of dealers in the city may not decline because the dealers will simply move to other markets.

In contrast, if the police crack down on the second market they might not be able to make it collapse. The enforcement pressure would push the wage down, making the market unattractive to mobile dealers, but dealing at this reduced wage might still be preferable to any other option available to immobile dealers. If so, then none of them would leave and the police would not be able to create a positive feedback effect until they applied considerably more pressure than would be required to collapse the first market.

If the police succeeded in collapsing the second market they would have accomplished something; they would have reduced the total number of dealers in the city because, by assumption, if immobile dealers cannot deal in their own market, they will not deal at all.

Suppose instead the police crack down on the third market. They would benefit from the mobile dealers' high reservation wage and positive feedback, so initially the number of dealers would decline at the same rate it would have in the first market. Suppose as much pressure was applied as was necessary to collapse the first market. Then all the mobile dealers would exit, leaving the immobile dealers. Then the market would look like the second market, only smaller. Perhaps enough smaller that the wage would fall below the immobile dealers' reservation wage, and the market would disappear completely forcing the immobile dealers to stop dealing.

If $E_{max}{}^{i}$ represents the effort needed to collapse market i and α is the fraction of immobile dealers in market 3, then assuming the three markets initially have the same number of dealers,

$$E_{max}{}^{3} = \mathrm{Max}\{\, E_{max}{}^{1}, \alpha\, E_{max}{}^{2}\,\}. \tag{4.71}$$

Hence the mixed market might offer the best opportunity. While it may be more difficult to collapse than the first market, it is easier to collapse than the second, and unlike the first, yields a reduction in the total number of dealers in the city if it does collapse.

This section considered how heterogeneity in dealers' reservation wages would affect the model, but there are other forms of heterogeneity. For example, some dealers are more violent than others. Violent dealers might congregate in one market, deter entry by less violent dealers, and command a higher than average wage. Studying this and other forms of heterogeneity would be a useful extension of the current work.

4.10 Estimating the Demand Parameter β

As was revealed above, the demand parameter β plays a key role in the model, so one must ask how it might be estimated. Two possible ways are to measure the elasticity of market size with respect to demand and the elasticity of demand with respect to the number of dealers. In symbols, these quantities are

$$\frac{\%\Delta N_{max}}{\%\Delta\alpha} = \frac{d\,N_{max}}{d\,\alpha}\,\frac{\alpha}{N_{max}} = \frac{1}{1-\beta} \qquad (4.72)$$

and

$$\frac{\%\Delta Q}{\%\Delta N} = \frac{dQ}{dN}\,\frac{N}{Q} = \beta. \qquad (4.73)$$

Note, the first result is the same whether one measures market size in terms of dealers or number of sales because

$$\frac{\%\Delta Q_{max}}{\%\Delta\alpha} = \frac{d\,Q_{max}}{d\,\alpha}\,\frac{\alpha}{Q_{max}} = \frac{1}{1-\beta}. \qquad (4.74)$$

Neither of these elasticities can be measured empirically because the independent variable is not controllable (or even easy to measure, particularly in the first case). The first is also difficult to estimate subjectively because it is directly affected by easing the positive feedback effect, and systems with feedback are difficult to understand intuitively.

Someone with first-hand knowledge of the market in question (for example, that neighborhood's patrol officer) might, however, be able to guess at the answer to the question, "By what fraction would sales increase if the number of dealers increased by 10%?"[61] That answer would give directly an estimate of β.

A little reflection might reveal which of two markets has the larger β even if neither value can be measured. For example, compare a market on a dead-end or little travelled street with one on a street that leads to other markets. The former probably has the larger β because customers will only visit it if they think there are dealers there.. In contrast, customers will travel the second street even if there are no dealers out.

Similarly, a market in which most of the dealers carry beepers may have a larger β if the dealers instruct their customers to come to that street to make their purchase. The more dealers there are, the more customers will come to that street.

It may, however, be that β does not vary a lot from market to market. It may vary instead over time. For instance, it might be larger when some wholesale suppliers have been arrested because fewer retailer dealers will be able to obtain drugs, so sales might be proportional to the number of dealers who were able to find an alternate supply.

4.11 Balancing Effort Against Users and Dealers

A common drug policy debate revolves around the question of what fraction of resources should be devoted to demand reduction and what fraction should be devoted to arresting and incarcerating dealers. The model above can by no means definitively answer this vital but difficult question, but it provides a framework for thinking about one small piece of it.

Suppose a city has decided it wants to clean up one of many open-air drug markets within its jurisdiction. It is considering cracking down on dealers in the manner described above and/or taking steps to reduce demand in that particular market. The city planners want to choose the mix of demand reduction and dealer enforcement that minimizes the effort required to eliminate the market.

Suppose that by expending $E_D(f)$ resources they can reduce the value of the demand proportionality constant α by 100f%. They might do this by publicizing plans to arrest dealers in that neighborhood, by arresting users in that market, by sending uniformed patrols through the market (presumably uniformed

[61] Of course to be precise one would ask about infinitesimal changes, but that might only confuse someone who is not accustomed to thinking in those terms.

patrols are relatively ineffective at capturing dealers because lookouts would warn the dealers, but the visible police presence could encourage prospective customers to go elsewhere), and/or by modifying traffic patterns to reduce the flow of traffic on the street (for example by changing the timing of stop-lights or changing one-way streets to two-way or vice versa). Let the units of $E_D(f)$ be such that applying one unit of effort against the dealers costs one unit. That is, normalize the measure of cost so that the total cost of demand and supply efforts is $E_D(f) + E$ where E is the enforcement variable in the model above.

The problem is to minimize $E_D(f) + E$ subject to the constraint that E be at least E_{max}, with α in the expression for E_{max} replaced by $(1 - f)\,\alpha$:

$$\underset{0 \le f \le 1}{\text{Min}}\ z(f) = E_D(f) + \left(\frac{1-\beta}{\gamma+\beta-1}\right)^{1/\gamma}\left(\frac{\gamma+\beta-1}{\gamma}\right)^{1/(1-\beta)} w_0^{1/\gamma}\, N_{max}$$

$$= \underset{0 \le f \le 1}{\text{Min}}\ z(f) = E_D(f) + \left(\frac{1-\beta}{\gamma+\beta-1}\right)^{1/\gamma}\left(\frac{\gamma+\beta-1}{\gamma}\right)^{1/(1-\beta)} w_0^{1/\gamma}\left(\frac{\pi\,(1-f)\,\alpha}{w_0}\right)^{1/(1-\beta)}. \tag{4.75}$$

It is difficult to even speculate about the nature of the function $E_D(f)$. It probably would vary from city to city, and perhaps even from market to market within a city. But just to complete the illustration, suppose $E_D(f)$ were linear in f, so that $E_D(f) = c\, f\, Q_{max}$ for some constant $c > 0$. Then the solution is

$$f^* = \text{MAX}\left\{1 - \left(\frac{(1-\beta)\,c\,Q_{max}}{E_{max}}\right)^{\frac{1-\beta}{\beta}},\ 0\right\} \tag{4.76}$$

Note that $c\,Q_{max}$ is the cost of eliminating the market using only demand reduction, and E_{max} is the cost using only enforcement directed at dealers. Call the ratio of these two expressions r. That is, define

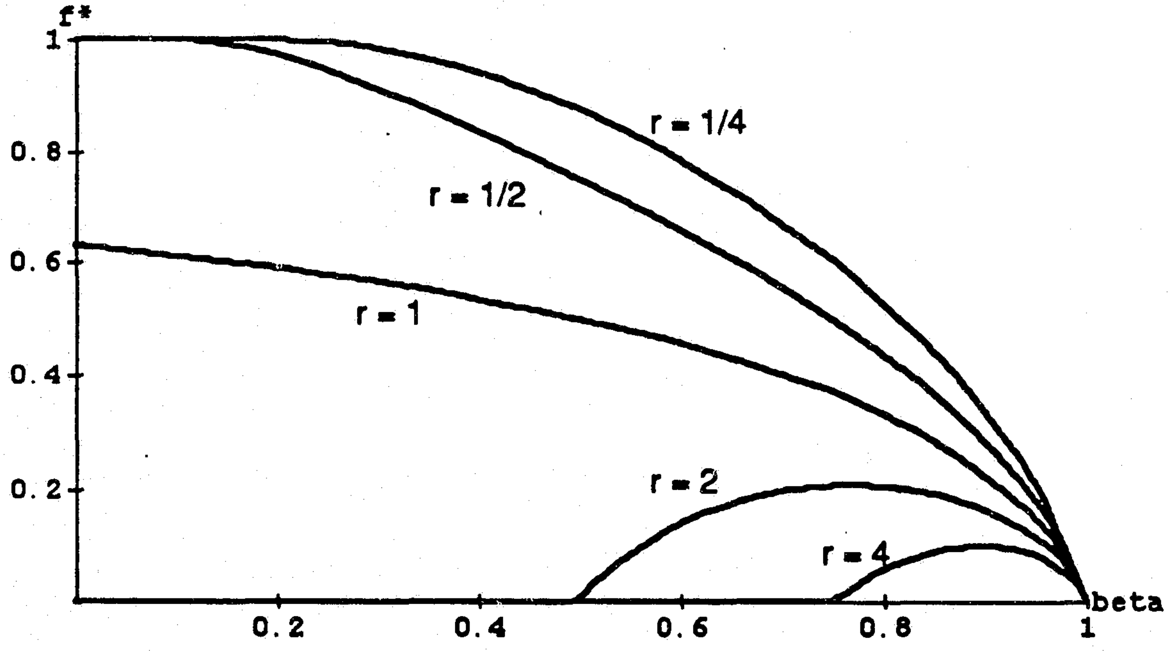
$$r \equiv \frac{c\,Q_{max}}{E_{max}}. \tag{4.77}$$

Then Equation 4.76 can be rewritten as

$$f^* = \text{MAX}\left\{1 - ((1-\beta)\, r)^{\frac{1-\beta}{\beta}},\ 0\right\}$$

Figure 4.17 plots f* as a function of β for various values of r.

Figure 4.17:
The Optimal Level of Demand Reduction
as a Function of β for Various r



Not surprisingly the smaller r is, and hence the less expensive demand reduction is relative to enforcement, the more the optimal policy relies on demand reduction. What is interesting is that if r is less than one, then the smaller β is, the more the optimal policy relies on demand reduction. This makes sense because when β is small, there is a surplus of dealers, so arresting dealers is relatively ineffectual. This simple relationship breaks down when demand reduction efforts get more expensive, but then demand reduction plays a relatively small role no matter what β is.

Figure 4.18:
The Optimal Level of Demand Reduction
as a Function of r for Various β

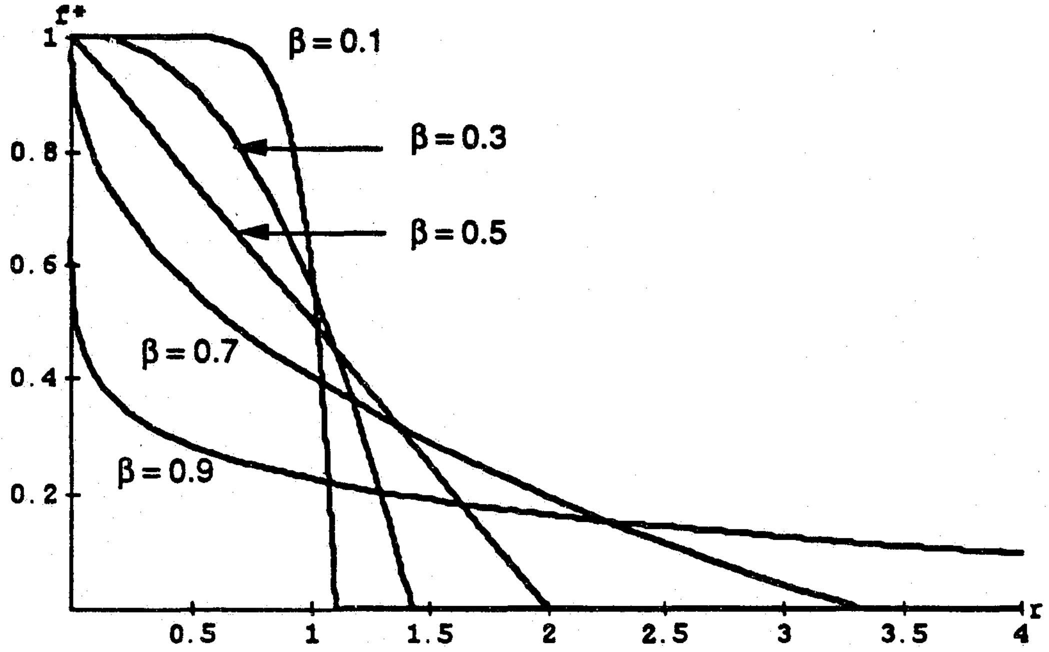


Figure 4.18 plots f* as a function of r for various values of β. It too shows something interesting. Again if β is large, f* is relatively small. Enforcement against dealers works best when β is close to one and dealers are in short supply. But for a wide range of r, some mixture of demand reduction and enforcement is optimal.

In contrast, when β is small, the transition region between relying primarily on demand reduction and relying primarily on enforcement becomes narrow. If β is as small as 0.1, the result approaches a bang-bang solution. If r is less than 1, the optimal solution relies almost exclusively on demand reduction, but if r is even slightly larger than 1, it is optimal to rely almost exclusively on enforcement against dealers.

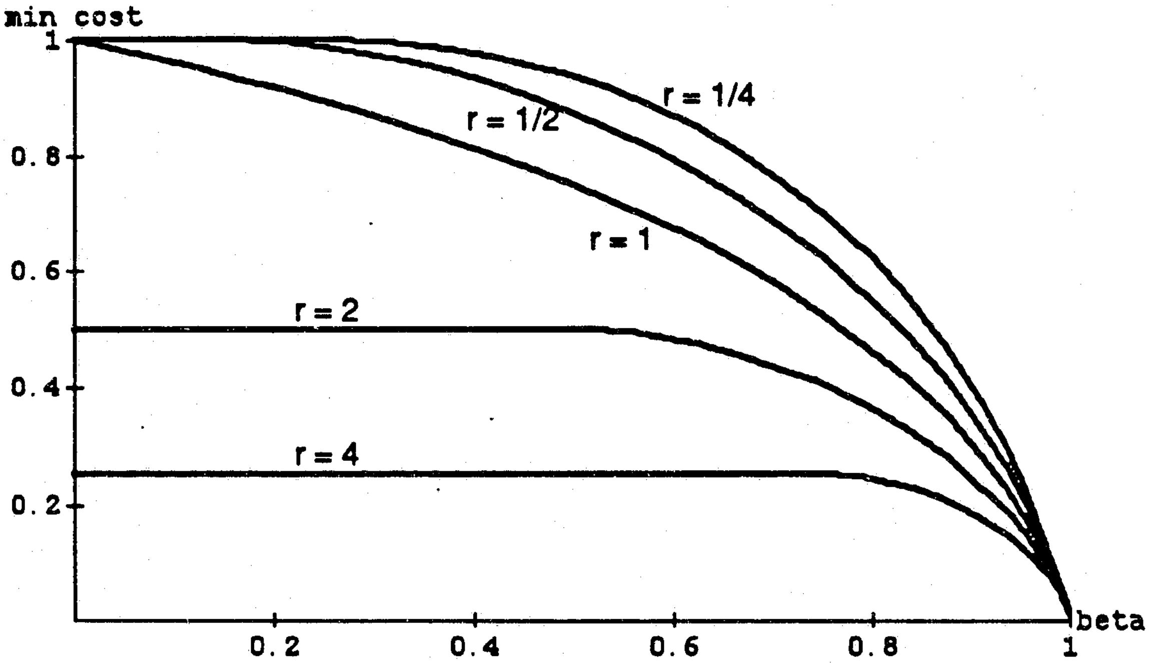
The minimum cost of eliminating the market is

$$c\, Q_{max}\, f^* + E_{max}\, (1 - f^*)^{1/(1-\beta)}$$

$$= c\, Q_{max}\left(1 - \beta\left((1-\beta)\, r\right)^{(1-\beta)/\beta}\right). \qquad (4.78)$$

Figure 4.19 plots the total cost as a fraction of the cost if only demand reduction is used as a function of β for various r. It shows clearly that the larger β is, the more important it is to use a mixture of policies instead of demand reduction alone.

Figure 4.19:
Minimum Total Cost of Eliminating a Market
as a Function of β for Various r



All of these graphs suggest following the rule of thumb, "go for the weak link." If β is small and hence there is a surplus of dealers, stress demand reduction. If β is large and hence dealers are in short supply, arrest dealers.

There may not always be sufficient resources to eliminate the market even if the optimal mix of demand and supply measures are used. Policy makers might then be interested in minimizing the volume of sales.

Suppose there are E_0 resources available. So an enforcement pressure of $E = E_0$ could be applied if no resources are allocated to reducing demand. The volume of sales can be viewed as a function of the demand parameter α and the enforcement level E, both of which are in turn functions of the policy mix parameter f. Hence the problem is:

$$\operatorname*{Min}_{0 \le f \le 1} Q(\alpha(f), E(f)) \tag{4.79}$$

where

$$\alpha(f) = (1 - f)\,\alpha \quad \text{and} \tag{4.80}$$

$$E(f) = E_0 - c\, f\, Q_{max}. \tag{4.81}$$

In general this problem cannot be solved analytically because there are not closed form solutions for Q(E).

To summarize this section, the balloon model provides a framework for thinking about the problem of dividing resources between demand reduction and supply control. At present it is not at all clear how many resources need to be expended to achieve a given reduction in demand. For the sake of illustration this section assumed a particularly simple relationship between effort expended and the reduction in demand. With this simple form two major insights can be derived. The first is "go for the weak link." The second is, some mixture of demand and supply control efforts is probably optimal if β is large, but if β is small, the optimal strategy probably relies almost exclusively on one or the other.

4.12 Modeling the A-Team/B-Team Phenomenon

This section tries to demonstrate that the balloon model can help formalize ideas about local enforcement other than those for which it was developed. Most of the discussion above addressed the question of how to manage local enforcement. One can also step back and ask the broader question, is local enforcement worthwhile?

Many people think local enforcement is futile simply because there are more dealers and potential dealers than could or should be incarcerated. Others go a step farther and argue that local enforcement may actually be counterproductive.

One rationale offered for this view is the "A-Team/B-Team model."[62] It asks one to consider what happens when the local dealers on a given street corner (the A-Team) are arrested. Customers will continue to visit that corner, so other people (the B-Team) will usually take their place. These may be the A-Team's lieutenants, other people in the neighborhood, or strangers.

[62] Explained to the author by John Coleman, Head of the Drug Enforcement Administration's New England Field Division, personal communication.

Typically the arrested dealers are not incarcerated very long. Distressingly often, when they are released they immediately return to "their" street corner. Then one of three things happens. At best the B-Team retires and the local enforcement accomplished nothing. If the B-Team continues dealing then either one of the teams moves to a new corner or there is a turf war. The first results in more dealers; the second creates street violence. Neither is desirable. Hence the A-Team/B-Team model suggests that at best local enforcement is useless, and it may make matters worse.

This pessimistic view can be formalized with a variation of the balloon model developed above. The variation assumes that both drug use and drug dealing are addictive. That is, it assumes that once users become "hooked" they will be slow to reduce consumption even if it becomes harder to find a connection, and that once someone has begun to deal, they will wait out periods of low demand rather than giving up dealing altogether.

More specifically it assumes that when the number of sales Q is less than the equilibrium number (αN^β, where N is the number of dealers) then sales will increase. But if Q is greater than this number, sales will remain constant. Sales will not increase because there are not "enough" dealers, but they will not decrease (at least not fast enough to make the assumption invalid) because the users are addicted.

This can be described as

$$\frac{dQ}{dt} = \begin{cases} c_2\left(\alpha N^\beta - Q\right) & \text{if } \alpha N^\beta \geq Q \\ 0 & \text{if } \alpha N^\beta < Q \end{cases}. \qquad (4.82)$$

The differential equation for N is the same as above (Equation 4.2) except that it is assumed that E is 0. Crackdowns are modelled directly as the jailing of 100f% of the dealers, instead of modelling them indirectly as an increase in steady state enforcement pressure.

Also, it is assumed that the constant c_2 for Q's differential equation is much larger than c_1, the constant for N's differential equation. That is, when the number of dealers and sales are not balanced, the number of sales will adjust more quickly (unless of course sales are "too high", in which case sales will not decrease).

One way of viewing this last assumption is that dealing (or at least the profits obtained from dealing) is also addicting. When the market turns sour for the dealers (there are not "enough" customers) they are reluctant to quit.

Suppose that initially the market is in equilibrium with N_0 dealers and $Q_0 = \alpha N_0^\beta$ sales. Now consider what happens if police arrest 100f% of the dealers, taking them off the street. Profits for the remaining dealers rise, so the wage exceeds the reservation wage w_0. By assumption, the number of sales remains at αN_0^β, so new dealers move in until the wage is reduced to w_0.

Now suppose the dealers who were arrested are set free. Then there are $(1 + f) N_0$ dealers on the street. At first wages are quite low, but the number of sales quickly increases to $\alpha [(1 + f)N_0]^\beta$. After this wages will be higher but still less than w_0. Then dealers will exit until the wage rises again to w_0, but fewer exit than entered. So when equilibrium is restored, there will be more dealers and more sales than before. Specifically, the number of dealers and the volume of sales will both rise to $100(1 + f)^\beta$% of their original values.

Table 4.1:
A Model of the A-Team/B-Team Phenomenon

Step	Q	N_{active}	N_{jail}	wage
Original equilibrium				
0	αN_0^β	N_0	0	$w = w_0$
A-Team arrested				
1	αN_0^β	$(1-f)N_0$	$f N_0$	$w = \frac{1}{1-f} w_0 > w_0$
B-Team enters				
2	αN_0^β	N_0	$f N_0$	$w = w_0$
A-Team released				
3	αN_0^β	$(1+f)N_0$	0	$w = \frac{1}{1+f} w_0 < w_0$
Shortly thereafter				
4	$(1+f)^\beta \alpha N_0^\beta$	$(1+f)N_0$	0	$w = (1+f)^{\beta-1} w_0 < w_0$
New equilibrium				
5	$(1+f)^\beta \alpha N_0^\beta$	$(1+f)^\beta N_0$	0	w_0
Final equilibrium				
∞	$(1+f)^{\beta/(1-\beta)} \alpha N_0^\beta$	$(1+f)^{\beta/(1-\beta)} N_0$	0	w_0

If this cycle is repeated the number of dealers and sales will increase again, although not by as much. In the limit, as this cycle is repeated an infinite number of times, the number of dealers and the number of sales will increase to $100(1 + f)^{\beta/(1-\beta)}\%$ of their original values. Table 4.1 describes the steps in the process. It shows clearly the "ratchet effect" by which local enforcement efforts make the market progressively larger.

The main point of this section is not to argue that the A-Team/B-Team model is correct. The credibility of that phenomenon derives from the wisdom and experience of the one who proposed the idea, not from the coincidence that it fits easily into the modelling framework developed above. To put it another way, the reader should decide whether or not the basic story is plausible before getting to the first equation.

Rather it is hoped that this section illustrates the versatility and usefulness of the modelling framework developed in this chapter. It allows one to formalize the A-Team/B-Team model. Formalizing the model helps one identify the key assumptions. For example, the story hinges on the fact that if the numbers of users and dealers are not in equilibrium ($Q \neq \alpha N^{\beta}$), then whichever quantity is in short supply will adjust (upward) while the other quantity remains (relatively) constant. In other words, both dealing and using are addictive.

If neither were addictive then local enforcement could ratchet the market down in size instead of up. Temporarily removing some of the dealers would induce some users to exit. Then when the dealers were released they would be greeted by less demand, so some would retire, leaving fewer dealers and fewer users than there were originally. If either dealing or using were addictive but not both, then periodically incarcerating some fraction of the dealers would have no long-term effect on the size of the market.

Clearly one could reach this insight without formalizing the model, but sometimes the process of formalizing it forces one to think rigorously, thereby identifying the key assumptions.

Formalizing the model can also alleviate some concerns about the verbal model. For instance, one might reject the verbal model because it seems to suggest that the market would grow indefinitely as long as the police periodically arrest some fraction of the dealers, which is not plausible. The formal model, however, suggests the market asymptotically approaches a well-defined bound.

4.13 Enforcement Pressure and the Number of Dealers

The balloon model assumed that the risk law enforcement imposes on each dealer is equal to the enforcement effort expended divided by the number of dealers, i.e., that it is E/N. This implicitly assumes that the total cost enforcement imposes on all dealers is equal to the effort expended E. Actually the total damage done (denoted by D) may be a function both of the effort expended (E) and the number of dealers (N), and hence the risk enforcement imposes on each dealer might better be modeled as $D(E,N)/N$, not just E/N.

The distinction arises because presumably the more dealers there are, the easier it is to catch one. So the total damage done with a fixed amount of effort may be an increasing function of the number of dealers present, i.e.

$$\frac{dD(E,N)}{dN} > 0. \qquad (4.83)$$

Obviously this could reduce the positive feedback effect. Removing a dealer would still leave more enforcement effort per dealer, but the enforcement would be less efficient.

If the total damage done by enforcement can be reasonably modeled as $D(E,N) = E\,N^{\phi}$ for some $0 < \phi < 1$, then

$$\left(\frac{D(E,N)}{N}\right)^{\gamma} = \left(\frac{E\,N^{\phi}}{N}\right)^{\gamma} = \left(\frac{E^{1/(1-\phi)}}{N}\right)^{\gamma(1-\phi)}. \qquad (4.84)$$

Hence the results above would still be valid if one replaced γ by $(1-\phi)\gamma$ and E by $E^{1/(1-\phi)}$.

More generally one can imagine at least four plausible scenarios for how Condition 4.83 might affect the positive feedback effect. One is that it could simply dampen the positive feedback. The market might still shrink for a time and then suddenly collapse, but it might shrink more slowly and only collapse after it had been reduced to a smaller size than was required before (i.e., E_{max} would be larger and both n_{min} and q_{min} would be smaller).

A second possibility is that the market would never collapse. If reducing the number of dealers made the remaining dealers better off, the market would never collapse. Without Condition 4.83 this could only occur when sales were constant ($\beta = 0$), but if D(E,N) were increasing in N, it could also occur with larger values of β.

Suppose that D(E,N)'s dependence on N did not become pronounced until the number of dealers has fallen by a certain fraction, specifically until the number of dealers fell below N_{min}. Then it might have no perceptible effect because the strength of the positive feedback is so great when the market is actually collapsing.

If it were able to halt the collapse, however, then the market would have two stable equilibria for many different levels of enforcement. In the high-volume equilibria enforcement's effect on an individual dealer is mitigated by dilution (safety in numbers). In the low-volume equilibria enforcement's effect on an individual dealer is mitigated by enforcement's ineffectiveness when there are few targets.

Observing some markets during crackdowns may be the best way to determine which of these scenarios holds. For now all that can be said is that if enforcement's effect is an increasing function of N then the positive feedback effect will be weakened. In the extreme case this might essentially negate the principal argument in favor of focused crackdowns, but it is also conceivable that it would have a relatively minor effect.

4.14 Summary of Results of the Balloon Model

This chapter developed a formal mathematical model that captures the spirit of the balloon metaphor. In doing so, it answers at least partially some of the questions raised in Sections 4.2 and 4.3. Of course the fact that the model suggests something does not make it right; all of these recommendations are tempered by the knowledge that the model is an abstraction and its assumptions are never fully satisfied.

4.14.1 Answers to Questions Raised in Section 4.3

According to the balloon model, if all of the model's assumptions are satisfied, the answers to the questions raised in Section 4.3 are:

(1) Is there any advantage to focusing effort on one market?

Yes. Except for the extreme case in which the number of sales is independent of the number of dealers ($\beta = 0$), there is positive feedback. That is, the incremental impact of an additional unit of enforcement pressure increases with enforcement pressure (until the market collapses).

(2) How hard does one have to push down to dent the market?

Equation 4.48 relates the steady state number of dealers in a dented market to the enforcement pressure E.

(3) If one pushes down hard enough, will the market pop (collapse)?

Yes. Equation 4.37 for E_{max} tells how much pressure is "enough".

(4) If so, how much will it gradually deflate before it pops?

Depending on whether one is interested in the number of dealers or the volume of sales, Equations 4.41 or 4.42 for N_{min} or Q_{min} respectively, give the answer.

(5) If one pushes down hard enough to partially deflate the market, but not hard enough to pop it, will the market spring back?

Yes, as is discussed in the introduction to Section 4.6.

(6) When a market is partially deflated or completely burst, is the dealing simply displaced to other markets or is it truly eliminated?

The balloon model does not answer this.

(7) If it is displaced, does it move only to adjacent markets or is it spread more or less uniformly over all the other markets?

Again, no answer.

(8) How much pressure is needed to keep a popped market from springing back?

As Section 4.8 explains, that depends on the number of dealers who try to "jump start" the market.

(9) Is the effort required to pop a market proportional to its size? To the square of its size? To some other power of its size?

Equations 4.37 and 4.39 show the effort required is proportional to the size of the market.

(10) What affects the proportionality constant?

According to this model, the proportionality constant is particularly sensitive to the demand parameter β, as Figure 4.9, Figure 4.10, and the discussion around Equation 4.40 show.

4.14.2 Answers to Questions Raised by Hartford's Plans

This section describes answers the balloon model would give to the questions raised in Section 4.2 about Hartford's plans assuming that all the appropriate assumptions are satisfied.

(1) How many and which of the 22 markets should be attacked first?

The crackdown should target only one market at a time, moving to another market only after the first has collapsed or it has been determined that there are insufficient resources available to collapse that market.

Hartford should only attack a market that it can collapse. If it cannot collapse the market in consideration, it should not even begin the crackdown; merely denting the market does not produce lasting results.

Assuming the police want to maximize the chance of collapsing a market, they should choose a market for which E_{max}, as given by Equation 4.37, is small. Since the effort required is proportional to the market size, this means they should choose a small market. And, among markets of a given size, they should choose the one for which the value of the demand parameter β is largest.

Finally, and perhaps most importantly, the police should attack a market which is unlikely to spring back if they do make it collapse. Which leads to the next question.

(2) How much pressure should be maintained on markets that have already been cleaned up when the main thrust goes on to other markets?

If there are no gangs or other mechanisms that might serve to coordinate dealers' actions, a relatively modest amount of maintenance pressure is required. If gangs are active in that area, considerably more pressure must be maintained.

Also, the smaller β is, the more pressure must be maintained to prevent the market from springing back.

Realistically, the amount of effort the police must expend to keep the market from coming back will be largely a function of how cooperative the citizens are.

(3) When should the crackdown begin?

Since the effort needed to collapse the market is proportional to its size, the crackdown should begin when the market is already smaller than it usually is. That is, the crackdown should begin when the demand parameter α is small, the profitability per transaction π is small, and/or the reservation wage w_0 is large.

The demand parameter α is probably smallest in the middle to end of the month,[63] during the week (and not over the weekend), and in bad weather. The profit per sale π is probably relatively constant, unless it decreases when there is an outburst of dealer-dealer violence. Cracking down during an episode of such violence might also increase the chances of obtaining community support. It may be that w_0 is largest during the school year because the younger dealers have something to do besides dealing. Obviously w_0 will be higher during good economic times, but α might also increase when the economy is strong, and waiting for a change in the nation's economy before beginning a local crackdown might be difficult to explain to the citizenry.

4.14.3 General Insights Derived from the Model

The model suggests a number of other insights which are described here.

(1) Go for the weak link.

It was argued that cracking down on dealers works best when dealers are in short supply (β is small).

Also, the police should not crack down with the objective of collapsing the market when the dealing is "fast and furious." The police may want to attack such a market for other reasons, for example, if their goal is to make as many arrests as possible. If the objective is to collapse a market, however, they should begin to crack down when the market is relatively quiet.

(2) Short, sharp crackdowns are mistakes.

The police should not make so many arrests in one day that they cannot maintain the pressure tomorrow. Markets can only be collapsed if pressure is maintained long enough for dealers to exit.

(3) Only begin a crackdown if you can finish it.

Denting a market does not permanently affect dealing if the market bounces back, and as was discussed in Section 4.1.4, crackdowns have negative side-effects. So police should only begin a crackdown if there is a reasonable chance they can collapse the market.

[63] Welfare checks come out early in the month, and they increase demand (personal communications with various members of the Hartford Police Department).

Furthermore, if the police cannot execute the crackdown without alienating the citizenry, thereby reducing the chance the neighborhood will resist attempts to restart the market, the crackdown should not be initiated.

(4) Gangs may play a key role in the formation of open-air markets.

4.15 Extrapolating Conclusions to Larger Markets

The balloon model was developed with local markets in mind, but it may apply to larger markets as well. The principal assumptions of the model are that dealers will enter if they can earn more than their opportunity wage w_0 and that the volume of sales is related to the number of dealers through $Q = \alpha N^{\beta}$.

At the level discussed above, w_0 was the wage available in other nearby markets. To apply the model to the national market, w_0 must be interpreted as the wage available in (licit or illicit) careers other than drug dealing. With that exception, the explanation of Equation 4.1 above applies to the national market.

Assuming that dealers' utility function has the form described above was a heroic assumption, but it is not much more heroic at the national level than it was at the local level. Dealers are dealers. Whether one thinks of them as participants in the national or local market, they are still the same people.

If anything some of the assumptions are less troubling at the national level. One might argue that if the national drug market grows then w_0 will rise because the criminal justice system will be able to devote fewer resources to apprehending and punishing non-drug offenders, and thus non-drug criminal careers become more appealing. However, this is likely to be a second-order effect. To first order, the unemployment level, the minimum wage, and other factors influencing opportunities in other sectors of the economy are probably not appreciably affected by the size of the drug trade.

The case for π being independent of N is a little harder to make. Above it was reasonable to make that assumption because changes in dealing on one street in one city are unlikely to affect the retail price or the price dealers pay. Changes in the size of the national market, on the other hand, might affect the profitability per transaction.

For example, if U.S. consumption grows substantially the import price might rise. Actually, this is probably not a significant effect. In the long term, which probably is not all that long, the international-level supply curve is fairly flat because there are no obvious limits on any of the factors of production.[64] In particular, it does not take much land to grow enough drug crops to satisfy demand, and there is no global shortage of farmers willing to supply the requisite labor. The recent decline in cocaine prices despite substantial increases in consumption is evidence of this.

Instead it is competition that might be more likely to affect π. As the market grows, each participant is likely to know more other participants, so it may become more difficult for dealers to maintain markups as high as they have in the past.[65]

On the other hand, economies of scale might reduce costs, and if retail prices are sticky, that could keep π from falling. Also, the very structure of the domestic distribution network might change if the market grew or shrank appreciably. Such changes could well affect π, although it is not clear in what direction.

At any rate, π may not be a constant when one examines the national market. However, for three reasons this need not keep one from at least gingerly exploring what the balloon model has to say about national markets.

First, it is not clear whether π is increasing or decreasing in the size of the market. When the direction of an effect is uncertain, it seems less likely that the magnitude of the effect is large.

Second, π appears in Equation 4.28 in the term $\pi \alpha N^{\beta-1}$. If π increases or decreases with N, that might be adjusted for by modifying β.

Finally, even if π depends somewhat on the size of the market, assuming π is constant may be a fair assumption for small changes in the size of the market.

All the discussion above is intended to suggest that Equation 4.29

$$\frac{dN}{dt} = c_1\left[\frac{\pi \alpha N^{\beta}}{N} - \left(\frac{E}{N}\right)^{\gamma} - w_0\right] \tag{4.29}$$

may be applicable at the national level.

Making the parallel argument for Equation 4.12

$$Q = \alpha N^{\beta} \qquad \beta \in [0,1] \tag{4.12}$$

[64] Moore, 1986.

[65] This possibility is discussed in Chapter 7.

is simpler. As before, sales volume is almost certainly increasing in the number of dealers and increasing at a decreasing rate (concave). Also as before, it is hard to say much more than that. Since Equation 4.12 fits these criteria, is at least plausible, and is convenient analytically, it seems as reasonable a guess as any other form.

The value of β may be different at the national and local levels, however. When the market is one of many open air markets in a city, the volume of sales may be appreciably affected by the number of dealers simply because mobile customers will naturally go to markets with lots of dealers. The more dealers there are the more likely there will be one ready to deal at any given point in time and the less likely it is that the dealer selected will be working with or under the observation of the police.

In contrast, national consumption would probably not be greatly affected by modest changes in the number of dealers. Various surveys indicate that drugs are already widely available to most people who might consider using[66] and the image of dealers as "pushers" who cajole novices into using is no longer widely held.[67] Hence, at the national level β is probably small.

For several reasons this strongly suggests that cracking down on the national market with enforcement oriented programs will not succeed in collapsing the market. First of all, the amount of pressure required to collapse the market is large when β is small and when the market is large. Second, if β is small, n_{min} is small, so one would expect to drive many dealers out of business before the market collapses. However, it does not appear that the number of dealers has been decreasing. That suggests that even after the massive increases in enforcement witnessed in the 1980's, the current levels of enforcement are still far short of those required to collapse the market. Finally, if β is small, then consumption responds to enforcement even less than the number of dealers does.

So the balloon model suggests there is essentially no hope that cracking down on the national market will make it collapse, and that denting the national market will not affect consumption appreciably. Furthermore, because of the positive feedback effect, enforcement is least cost effective at lower levels of intensity. Hence pressing down uniformly over the entire national market is probably particularly inefficient.

The balloon model may have another important implication for the national market if β is small. Recall the discussion in Section 4.11

[66] U.S. Department of Health and Human Services, 1988c, pp.153-157.

[67] See, for example, Kaplan (1983a).

that if β is small then the optimal mix of demand reduction and supply control will include almost all of one and none of the other. This suggests that the popular notion of dividing resources equally between demand reduction and supply control may not be optimal.

This observation is tentative at best for at least three reasons. First, the analysis in Section 4.11 rested on a particular, simple relation between the costs and benefits of demand reduction. Second, that result gave the minimum cost way to collapse the market. Collapsing the national market probably is not feasible, so it may be more appropriate to ask how can consumption be minimized short of collapsing the market. Third, a 50/50 mix might well be at least approximately optimal if the criterion is minimizing the expected total cost, with the expectation taken over the ratio of the total cost for an exclusively demand oriented program to the total cost of an exclusively enforcement oriented program.

One alternative is to target particular cities. Recent national drug strategies have place special emphasis on Washington, D.C.,[68] although some are already pronouncing these efforts to be a failure.[69] Perhaps the target was too large. The balloon model suggests that more could be accomplished by focusing on smaller targets and applying enough pressure to collapse them.

Giving some cities special attention might be politically difficult, however. What representative would be willing to allow the bulk of federal drug enforcement resources to be allocated to targets outside his or her district?

Another way to avoid spreading federal resources uniformly over the national market, and hence diluting them to the point of uselessness, would be to focus on something other than a geographic target. For example, intense enforcement attention may have successfully limited the mafia's role in drug dealing. Today a comparable target might be Jamaican posses. Their unusual level of violence may warrant such attention.

Or the focus could be on a particular drug. The cocaine market might be simply too large already for there to be much hope of achieving some positive feedback at the national level. That might not be the case for heroin,[70] however, particularly in view of the fact that AIDS may independently reduce the size of that market (See Chapter 6).

[68] Berke, 1989.

[69] Miller, 1990.

[70] Reuter and Kleiman (1986) argue that enforcement may be most effective against heroin.

If there is little hope of collapsing the national market through enforcement, then there is no point in exploring quantities such as N_{min}, Q_{min}, and E_{max} at the national level. The expressions for N_{max} and Q_{max}, however, may yield some insight.

Quantities such as π, w_0, and to a lesser extent α are largely beyond the control of local police, so when the balloon model was applied above, there was little discussion of the policy implications of changing those parameters. That need not be the case at the national level.

For example, a concerted campaign against demand (either by enforcement directed at users or through an education campaign) could reduce α. Equations 4.32 and 4.33 suggest that reducing α would have proportionate effects on the number of dealers and the volume of sales, and that if β is indeed small, the changes would be of the same order of magnitude as the change in α.

Changing w_0 would have a very different effect. Some people advocate redirecting resources spent on enforcement to jobs and anti-poverty programs for the inner city. Such programs can be viewed as increasing w_0, the appeal of the dealers' best alternative to selling drugs.

Equation 4.32 suggests that increasing w_0 would in fact decrease the number of dealers. If β is small, then for small changes in w_0, the corresponding change in N_{max} would be of the same magnitude. For example, increasing w_0 by 20% would decrease the number of dealers by about 20%.

Equation 4.33 suggests, however, that if β is small, increasing w_0 would have much less impact on the volume of sales. For instance, if $\beta = 0.1$ then even doubling w_0 would only reduce sales by about 7.5%. The reason for this is simple. If β is small, dealers, and hence drugs, are not in short supply. People who want to buy will be able to continue to buy even if the number of dealers is substantially reduced.

This does not mean the government should not fight poverty. It just suggests that people should not expect anti-poverty programs to appreciably affect the availability of drugs.

In summary, applying the balloon model to the national market suggests that demand reduction efforts are likely to be the most effective. "Cracking down" at the national level would almost certainly not be effective.

CHARACTERISTICS OF THE SUPPLY AND DEMAND OF ILLICIT DRUGS

Jonathan P. Caulkins

Chapter 7: Characteristics of the Supply and Demand for Illicit Drugs

7.0 Introduction

The preceding chapters explored particular issues surrounding the markets for illicit drugs. This chapter steps back and examines some unusual characteristics of the industry-level supply and demand for illicit drugs. Clearly there is not one market for illicit drugs, any more than there is one labor market or one capital market in the U.S. economy; the market can be subdivided by level, location, and type of drug. Nevertheless, there are times when a brief glance at the forest reveals things that are difficult to see even after carefully examining all the trees individually. For example, macroeconomists do speak of the labor market and the capital market as if there were only one because it is difficult to explain inflation and the business cycle using only the tools of microeconomics. Likewise, the phenomena discussed here are characteristics of the illicit drug industry as a whole, not of the individual participants. So this chapter will speak of the supply and the demand curves for illicit drugs even though this is a great simplification.

Section 7.1 looks at how the cost of providing drugs varies with the quantity consumed; i.e., it looks at the supply curve for illicit drugs. It is argued that the supply curve is downward sloping because the larger the market is, the more efficiently it operates and the lower the costs imposed by enforcement. A downward sloping supply curve allows for multiple stable market equilibria and hence could explain the phenomena of relatively low consumption at fairly high prices seen before 1970 and the high consumption at relatively low prices observed today.

Section 7.2 discusses some implications of this model including the suggestion that there are limits to what enforcement can accomplish given today's high consumption and low prices. Section 7.3 argues, however, that this does not mean legalization or even significant reductions in enforcement are necessarily good ideas.

Section 7.4 turns to demand. Several drug market researchers have suggested that the price elasticity of demand for drugs is probably relatively small in the short run, but larger in the long run. The explanation for this is that addicts' demand is relatively unresponsive to price, but when prices rise, fewer non-addicted users become addicted. Section 7.4 proposes a functional form for a demand curve that captures this effect and examines some of its implications.

7.1 A Static Model of Multiple Equilibria

7.1.1 Historical Evidence That Needs Explaining

A gross oversimplification of the post-WWII history of drug use in America is that for decades there was relatively little use. Then, in the late 60's and 70's there was an explosion in drug use. Now drug use is widespread, but with conspicuous exceptions for certain drugs and certain demographic groups, drug use appears to have stabilized somewhat, although at a vastly greater level than before. How, one might ask, could society have had two such radically different "stable" consumption patterns?

One answer is to deny the validity of this gross oversimplification of historical trends. One might dispute, for instance, the assertion that the pre-60's period of relatively low use was indeed a stable pattern of consumption. Perhaps it was never stable; perhaps it was like a powder keg waiting for a spark to set it off.

Likewise one could question whether the current situation is stable. Some would say that drug use is still growing; some even feel the rate of increase is itself still increasing. Their model of historical trends in drug use might be that of a nuclear chain reaction; once some critical mass is reached, drug use spreads through some unstoppable chain reaction.

Some observers, albeit a minority, are at least hopeful (if not expectant) that drug use will decline substantially. The principal cause for such optimistic projections is that the epidemic of drug use which swept the country in the late 19th and early 20th centuries subsided more or less of its own accord after reaching proportions comparable to those of the current epidemic.[1] These observers' model of trends in drug use might be like that of an isolated animal specie. Initially a population living in isolation grows exponentially, but when it exceeds the environment's carrying capacity and depletes food resources, the population comes crashing down. It is possible that what appears today to be a stable equilibrium with relatively high drug use may in fact be just the peak of one cycle in a history of booms and busts in drug consumption.

Finally, even if one believes there have in fact been two distinct periods of stable use at vastly different levels, the existence of two such equilibria need not be a conundrum. After all, they occurred several decades apart and both tastes and supply change. Perhaps by today's standards relatively few people before the 60's valued highly the experience of using drugs, so there was little consumption. Then interest in drug use grew, and as a result so did

[1] Musto (1987) describes this earlier drug epidemic.

consumption. Now, perhaps demand has ceased growing, so consumption has once again stabilized.

Today's higher consumption is accompanied by relatively low prices. If the supply curve is indeed downward sloping as will be argued, increased demand alone can explain the higher quantities and lower prices, and the increase need not even be permanent. If supply is not downward sloping, however, then the supply curve must have shifted out as well as the demand curve. This too could certainly have happened over the course of time.

7.1.2 An Explanation Based on Downward Sloping Supply

This section does not in any way attempt to refute the four explanations above, but it does offer an alternate view. It suggests that even in a static world in which parameters such as consumers' preferences, enforcement resources, the risks and costs of drug dealing, and so on are all fixed and unchanging, it is possible to have two stable equilibria, one at relatively low levels and one at high levels of consumption.

The key to this explanation is that the supply curve is downward sloping; the larger the quantity of drugs supplied, the lower the per-unit cost of supplying those drugs would be.[2] No special properties are assumed of the demand curve.

Consumers buy goods, including drugs, until the marginal utility derived from the drugs is offset by the marginal cost of obtaining them. Roughly speaking there are three categories of costs: costs imposed by enforcement against users, search time costs, and the dollar price of drugs. Clearly these costs are interrelated, but distinguishing them facilitates the discussion. The next few paragraphs analyze how these costs depend on the size of the market, measured in terms of the quantity of drugs consumed.

The dollar price of drugs as a function of quantity is just the industry supply curve. The supply curve describes the prices at which the drug distribution industry would be willing to provide various quantities of drugs.

In conventional industries the supply curve is drawn with an upward slope because there are generally fixed factors, for instance the physical plant.[3] As production increases, the industry uses more variable factor inputs, such as labor and raw materials. When the ratio of variable to fixed factors increases beyond some point, production becomes less efficient and the cost per unit rises.

[2] This idea was suggested to me by Mark A.R. Kleiman.

[3] See Varian (1984, Chapter 1) for a discussion of the conventional theory of the firm and supply curves.

Usually a distinction is made between the short run and the long run. In the short run more factors are fixed than in the long run, so the industry supply curve slopes up more steeply. Unless there are significant economies or diseconomies of scale, or the industry is so large it bids up the price of factor inputs, the long run supply curve is usually thought to be fairly flat.

Moore tries to identify fixed factors for the illicit drug industry and concludes that connections, trustworthy supplier-buyer relationships, may be the only significant, limiting factor.[4] The provision of drugs is labor intensive, and most of the labor is unskilled. Since there is a large surplus of unskilled labor willing to work at the wages offered by the drug industry, rapid expansion of industry capacity is possible.

The other raw material is the drugs themselves. Source country production capabilities are very large,[5] and can be expanded quickly.[6] And as Reuter, Crawford, and Cave[7] point out, smuggling resources, except perhaps for skilled pilots, are not in short supply. So, there do not appear to be significant limitations there.

Very little capital is required. Dealers are generally brokers. The most processing they usually do is diluting the drugs and repackaging them,[8] neither of which requires any significant machinery.[9] Even a synthetic drug laboratory requires far less capital than the term suggests.

So the supply curve for the illicit drug industry is not likely to slope up for the usual reasons, except perhaps in the very short run. There are other costs that contribute to the supply curve for illicit drugs, however, that are not like the factor inputs to a traditional industry. So the way they depend on the quantity consumed probably determines how the entire illicit drug industry supply curve depends on quantity.

[4] Moore, 1986.

[5] As pointed out in Chapter 2, world opium production is 30-50 times greater than what is required to supply the U.S. heroin market. The situation for cocaine and marijuana is not so extreme, but there is more than enough capacity to supply the U.S. market (Reuter et al., 1990, p.240).

[6] There have been several instances in the last twenty years in which production of a drug has been eliminated or at least greatly reduced in one country only to have production spring back shortly thereafter in another country. This happened, for example, when Colombia replaced Mexico as the principal source of marijuana consumed in this country.

[7] Reuter, Crawford, and Cave, 1988.

[8] Dealers also convert powder cocaine into crack, but that does not require special skills or equipment either.

[9] The most sophisticated machinery for some dealers may be their guns and a money counter.

Roughly speaking these other costs can be divided into three categories: the costs of enforcement, the cost of making connections, and costs arising because of the drugs' illegality (including the costs of robbery, dealer-dealer violence, security precautions, and so on). It will be assumed here that the costs per unit delivered belonging to the third category do not vary as the market grows.

When the market is small the risk of arrest for an individual dealer probably does not change much if the market grows or shrinks a little, and the punishment upon arrest is determined by statute, so it does not depend much at all on the size of the market. However, the criminal justice system's punishment resources are limited. The limit might be determined, for example, by the amount of prison space available. As the market grows beyond the point at which the statutory punishment would fill all the prison space society is willing to allocate to drug offenders, the amount of punishment per market participant begins to fall. Beyond that point, doubling the number of dealers will roughly halve the enforcement cost per dealer. Assuming the size of the market is proportional to the number of dealers, this would then halve the enforcement cost component of the industry supply curve.

The market might also become more efficient as it grows because the cost of making connections would decrease. Reducing the cost of making connections directly reduces the cost of providing drugs. Also, the more alternate supply sources dealers have, the more they can shop around and bid down prices. This increased competitive pressure would presumably reduce costs by squeezing out excess profits and eliminating inefficient business practices.

Thus both the enforcement costs and market efficiency arguments suggest the supply curve for illicit drugs may actually slope downward; the more drugs the industry supplies, the less expensive it is per-unit to supply them.

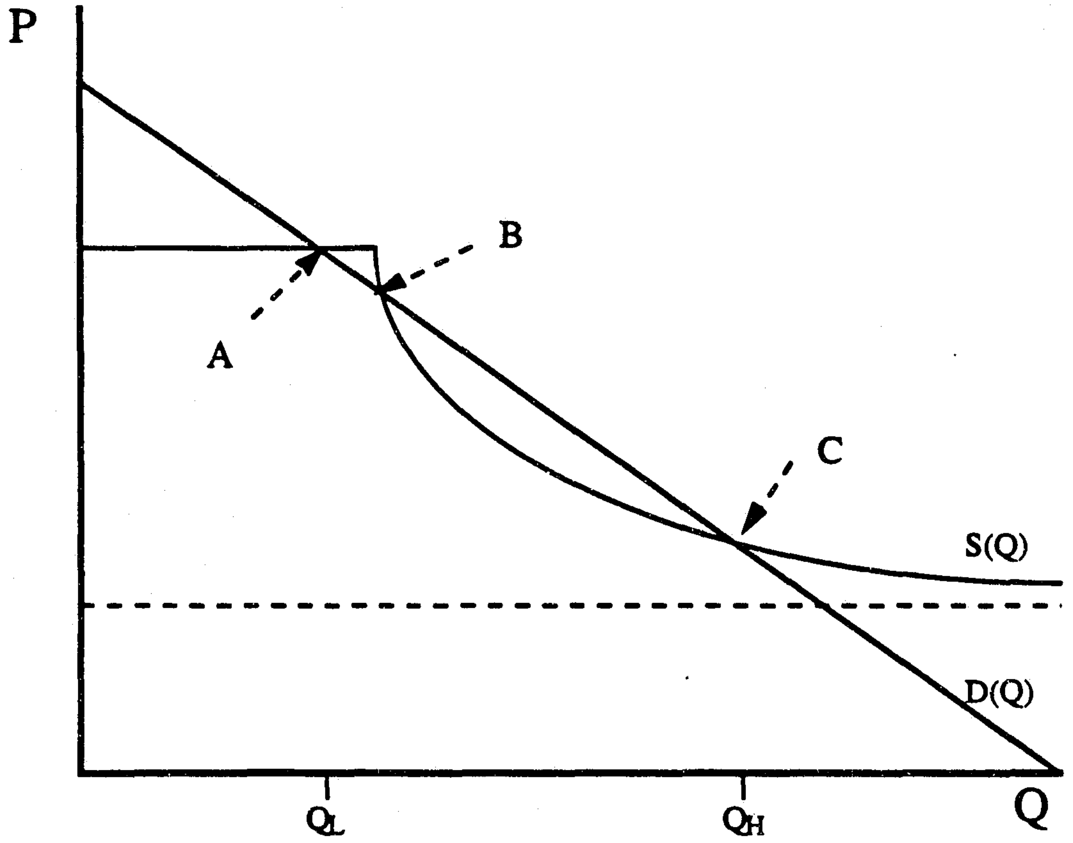
Recall that users face significant costs other than dollar costs and that these costs affect the equilibrium quantity consumed. These costs probably depend on market size in ways similar to the ways the dealers' costs that were just discussed do. The cost premium associated with enforcement against users is probably constant out to a certain market size, at which point the criminal justice system becomes saturated, and then falls after that. Likewise search time costs, which are analogous to the dealers' costs of making connections, might decline as the size of the market increases. The more buyers there are, the more dealers there will be, and the more dealers there are, the lower the search time costs will be.

These non-price costs to users affect the market equilibria. The proper way to analyze them would be to draw the industry supply curve and then another curve above it representing the dollar costs (given by the supply curve) plus the non-price costs.[10] The intersection of this second curve with the demand curve would determine the equilibrium quantity. The supply curve evaluated at that quantity would give the equilibrium dollar price.

However, drawing the extras curves clutters the diagrams. Since the non-price costs have the same general shape as the variable part of the supply curve, the conclusions of the qualitative analysis done below are not affected if one simply works with the industry supply curve.

Given the discussion above, the industry level supply and demand curves (denoted S(Q) and D(Q) respectively) might look like those depicted in Figure 7.1. The vertical axis gives the per unit price or cost. The horizontal axis gives the total quantity of drugs sold (the size of the market).

Figure 7.1:
A Downward Sloping Supply Curve That Gives Two Stable Equilibria



[10] This is how elementary economics texts analyze taxes and anything else that makes the seller's revenues differ from the buyer's costs.

The supply curve is roughly horizontal out to the point at which the criminal justice system becomes saturated. Thereafter the cost per unit of supplying drugs decreases as the size of the market increases. Costs do not decrease to zero; they approach the horizontal dashed line which represents per unit costs that do not depend on the size of the market. These include the cost of the dealers' own time, the cost of purchasing the drugs, the direct transportation and packaging costs, and the costs of robbery and dealer-dealer violence.

Point A is a stable equilibrium at a relatively low level of use (denoted Q_L for low quantity). It is an equilibrium because costs equal benefits, so the market clears. To see that it is a stable equilibrium consider the points around it, for instance a point to its right. At such a point the marginal users who derive the least satisfaction from using derive less benefit than it costs to supply the drugs they consume.[11] Since dealers will not sell below cost, the marginal users would exit and the market would return to point A. On the other hand, at points to the left of A there are people who are not currently using, but who would derive benefits exceeding the price at which dealers would be willing to sell to them. One would expect those individuals to begin using, moving the market to point A.

In general any intersection of the supply and demand curves gives a market equilibrium. If the slope of the supply curve is greater (less negative) than the slope of the demand curve, the equilibrium will be stable. Otherwise it will be an unstable equilibrium. Alternately, when the demand curve is above the supply curve, more mutually beneficial sales can be made, so the quantity sold will increase. But if the supply curve is above the demand curve, then the cost of supplying drugs to the marginal users exceeds the benefits they derive, so those users will exit the market and the quantity sold will decrease.

Hence point B is an equilibrium, but it is not stable. Point C is another stable equilibrium, but the level of consumption at point C (denoted Q_H for high quantity) is far greater than at point A and the per unit cost of supplying drugs is lower. The additional users all derive an intermediate amount of satisfaction from using. If the market were small (point A) such individuals would not use because the costs (principally of enforcement) outweigh the benefits. But if many people are already using (point C) these individuals would use as well. At point C the benefit they derive exceeds the fixed costs

[11] Of course the marginal consumption may be from someone who consumes a smaller but still positive amount when the price is lower. But for ease of explication the discussion is phrased as if there is an individual who will only consume at all if the price is below the price under consideration.

unrelated to enforcement. The per unit costs imposed by enforcement are small because enforcement is spread out over so many transactions, and the search-time/making-connections costs are smaller because the market is larger. Safety in numbers allows people with intermediate valuations to use if and only if many others are using. This negative externality is one of the ways that individual users, even so-called "casual" users, contribute to "the drug problem."

On the whole point C seems to describe the current high quantities and relatively low prices better than point A. Even today, however, the criminal justice system has not utilized its maximum capacity for punishing drug offenders. To be sure many jails and prisons, especially at the state and local level, are filled beyond their rated capacity, but more facilities are being built; it is conceivable that still more people could be packed into existing facilities; and there are sanctions, such as fines and community service, that do not involve incarceration. Furthermore, not every inmate currently in prison was incarcerated for a drug offense. So even if every time a new drug offender is sent to prison, someone else is released to make room, arresting and sentencing drug offenders increases the total enforcement cost against drug market participants because some of those released would have been serving time for other offenses. So the criminal justice system as a whole is not yet in the situation of simply redistributing a fixed, finite amount of punishment among drug offenders. In large cities it might be nearly that bad; in some smaller cities and towns the situation might be much closer to that described by point A.

7.2 Policy Implications of Multiple Equilibria Model

The model of a downward sloping supply curve and ensuing multiple stable market equilibria, assuming it has some validity, has important policy implications. These are discussed next.

7.2.1 Importance of Responding Quickly

If society was originally at point A and is now at point C, one might well ask how it moved from one to the other? One possibility is that demand shifted out temporarily, pushing the equilibrium quantity to the right of point B. Then even if demand shifted back later, the market would continue growing to point C. A second possibility is that both demand and enforcement effort have been growing over time, and enforcement effort may even have grown in proportion to demand, but with a lag. The lag could have allowed society to move from point A to point C.

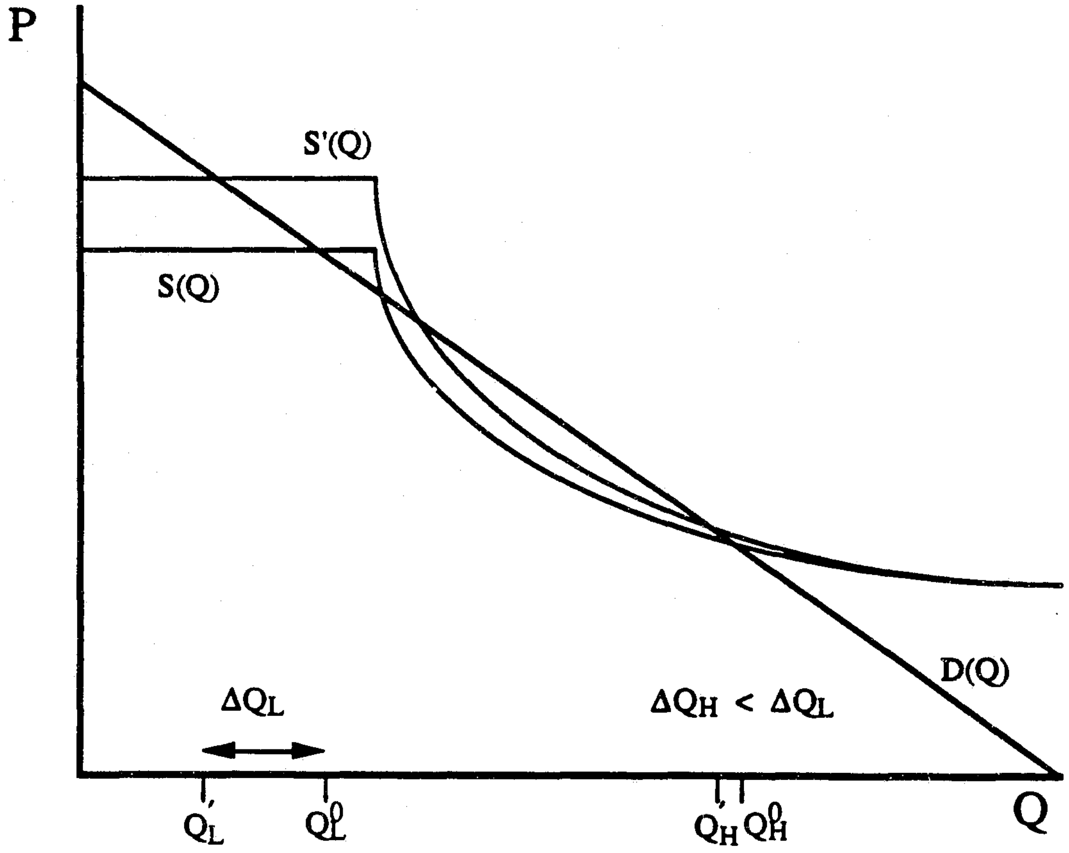
In either case there is an important lesson for communities that are still at point A. In the long run such communities would save considerably if they responded to increases in demand quickly, before the situation deteriorated from point A to point C.

7.2.2 The Effectiveness of Enforcement

Recent policy has stressed enforcement. The basic idea is that by increasing enforcement related costs, the government can shift the supply curve up and hence reduce consumption. The analysis in Section 7.1 suggests that this approach is far more likely to be effective if the market is at point A than if it is at point C.

Suppose that by increasing enforcement one means increasing the severity of punishment and/or the likelihood of being arrested and also increasing the criminal justice system's punishment resources by enough that it can handle the same size market before having to ration punishment. Then the supply curve would shift up from S(Q) to S'(Q) as depicted in Figure 7.2.

Figure 7.2:
The Effect on Increasing Enforcement



If the market were originally at point A then such a measure would work as expected. The enforcement cost per transaction increases, and as a result the quantity consumed decreases appreciably (by the amount ΔQ_L).

Suppose on the other hand the market were originally at point C. Then the intersection of the demand curve and the new supply curve is only slightly to the left of the intersection with the original supply curve, so the decline in consumption is small ($\Delta Q_H < \Delta Q_L$). This is simply because in the vicinity of point C enforcement costs are less important than other costs, and increasing something that is relatively unimportant, even if it increases substantially, will have a limited overall effect.

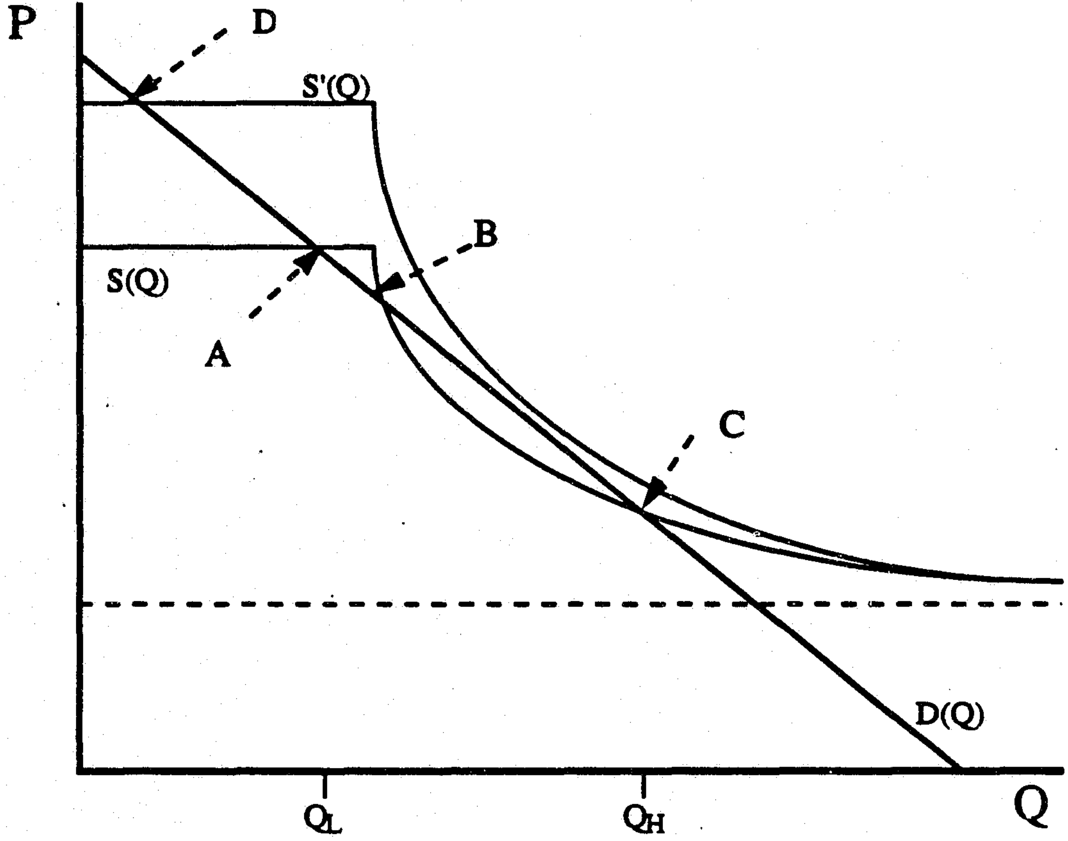
Hence, this model suggests that even if increasing the enforcement effort would be successful at point A, it may be relatively ineffectual if consumption has stabilized at the high levels described by point C.

7.2.3 Comparison with the Balloon Model

The discussion above suggests that consumption is relatively unresponsive to small increases in enforcement when the market is at point C. The model also suggests, however, that enforcement could work spectacularly in some cases. Suppose the government were temporarily able to marshall a massive enforcement effort and could shift the supply curve up from S(Q) to S'(Q) as depicted in Figure 7.3. (The demand curve is drawn with a steeper slope to make the point.)

Then the per unit cost of supplying drugs would exceed the benefit derived by the marginal users and some people would stop using. As they did, enforcement's contribution to the per unit costs would rise, further shrinking the market. This synergistic feedback would drive the market all the way back to point D. Then, even if enforcement were subsequently reduced to its original level, consumption would only move out to point A. So a massive, temporary crackdown could conceivably achieve substantial long-term reductions in drug use if it were maintained long enough to drive the market below point B. Hence the multiple equilibria model gives results similar to those obtained with the Balloon Model in Chapter 4.

Figure 7.3:
Possibility of Collapsing the Market With Enforcement

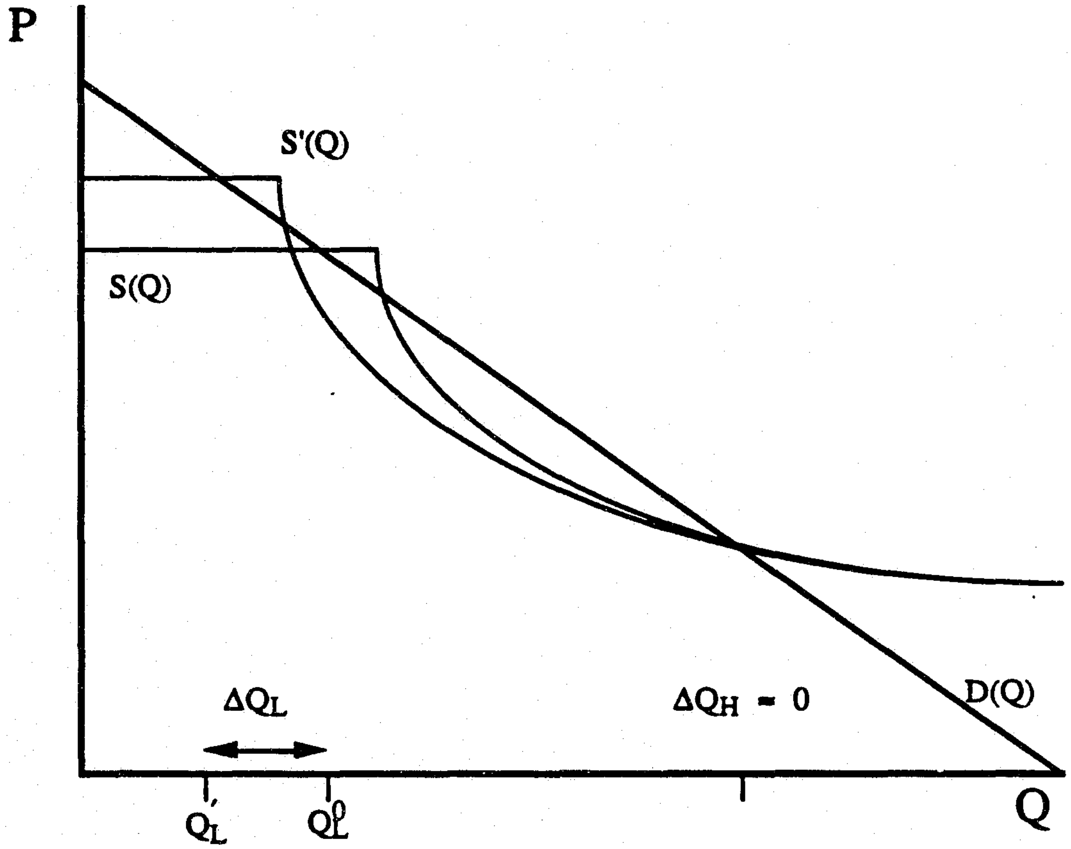


It is not clear how massive such an effort would have to be, nor how long it would have to last. The model is far too simplistic to even begin to shed light on such vital questions. Nevertheless it offers a glimmer of hope for enforcement. It is doubtful that such a massive crackdown could be achieved at the national level, but it may be feasible for small to medium-size cities. Naturally the resources available to such a city are proportionately smaller, but one can imagine gathering significant federal resources for a crackdown on drug use in one city, driving it from situation C to situation A, and then moving on to another city in the hope that, although the first city could not have driven consumption from C to A without assistance, it might be able to keep it stabilized at a relatively low level of use.

7.2.4 The Effect of Imposing Stiff Minimum Sentences

Next consider the effect of imposing harsh minimum sentences without increasing the criminal justice system's punishment capacity (prison space). This would increase the cost of using at the low quantity equilibrium (point A), but it would have no effect on the cost of using at point C where all available punishment capacity was already in use.[12] (See Figure 7.4.)

Figure 7.4:
The Effect of Imposing Stiff Minimum Sentences

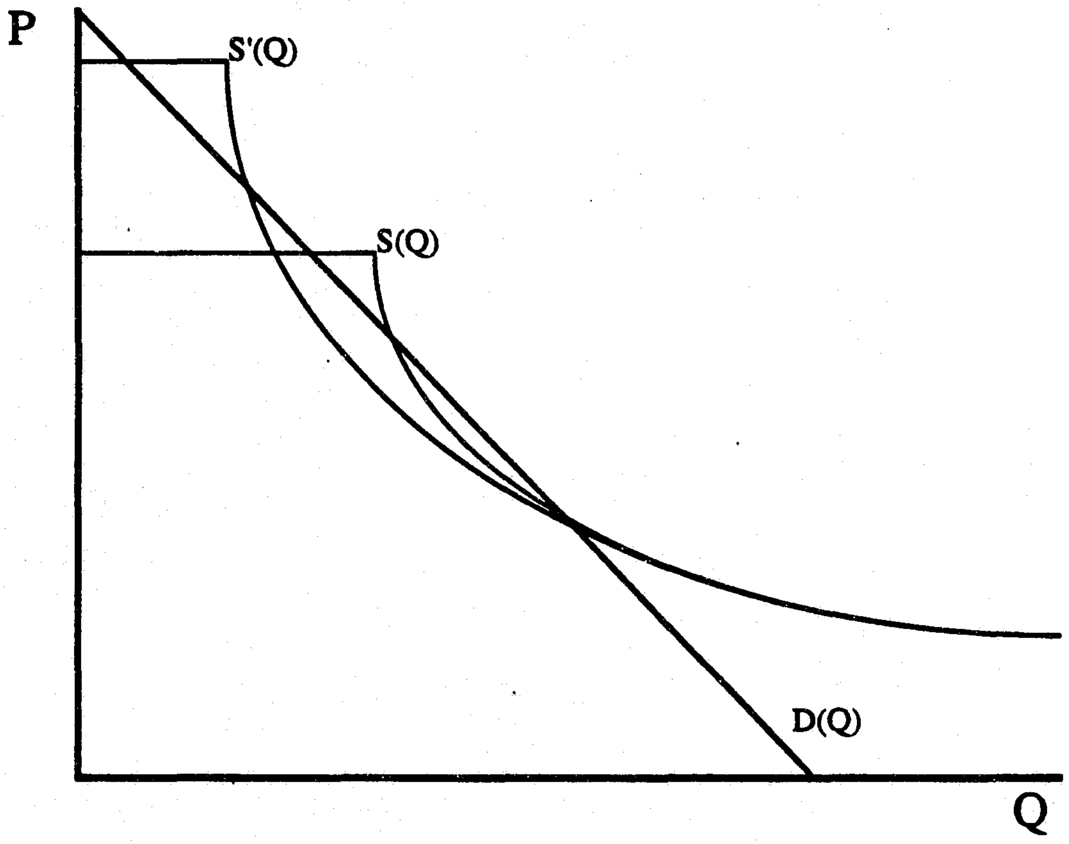


The other effect of minimum sentences would be to move to the left the kink in the supply curve representing the point at which the criminal justice system's punishment capacity becomes fully utilized. If the demand curve is sufficiently steep, the market were originally at point A, and point A were close to the kink, then it is possible that imposing harsh minimum sentences (shifting the supply curve from S(Q) to S'(Q)) could move point B to the left of the current market size. Then the demand would be above the supply curve and

[12] This assumes the cost of enforcement is proportional to the expected punishment. To the extent that market participants are risk averse, imposing fewer, longer sentences would increase the cost. However, it is conventional wisdom that a more certain punishment has a greater deterrent effect even if it is less severe. To the extent that deterrence and cost are related, stiff mandatory sentences might actually reduce the cost of acquiring drugs if the criminal justice system's punishment resources are already fully utilized.

equilibrium would not be restored until the market reached point C. (See Figure 7.5.)

Figure 7.5:
The Possibility That Stiff Minimum Sentences
Will Lead to Greater Consumption



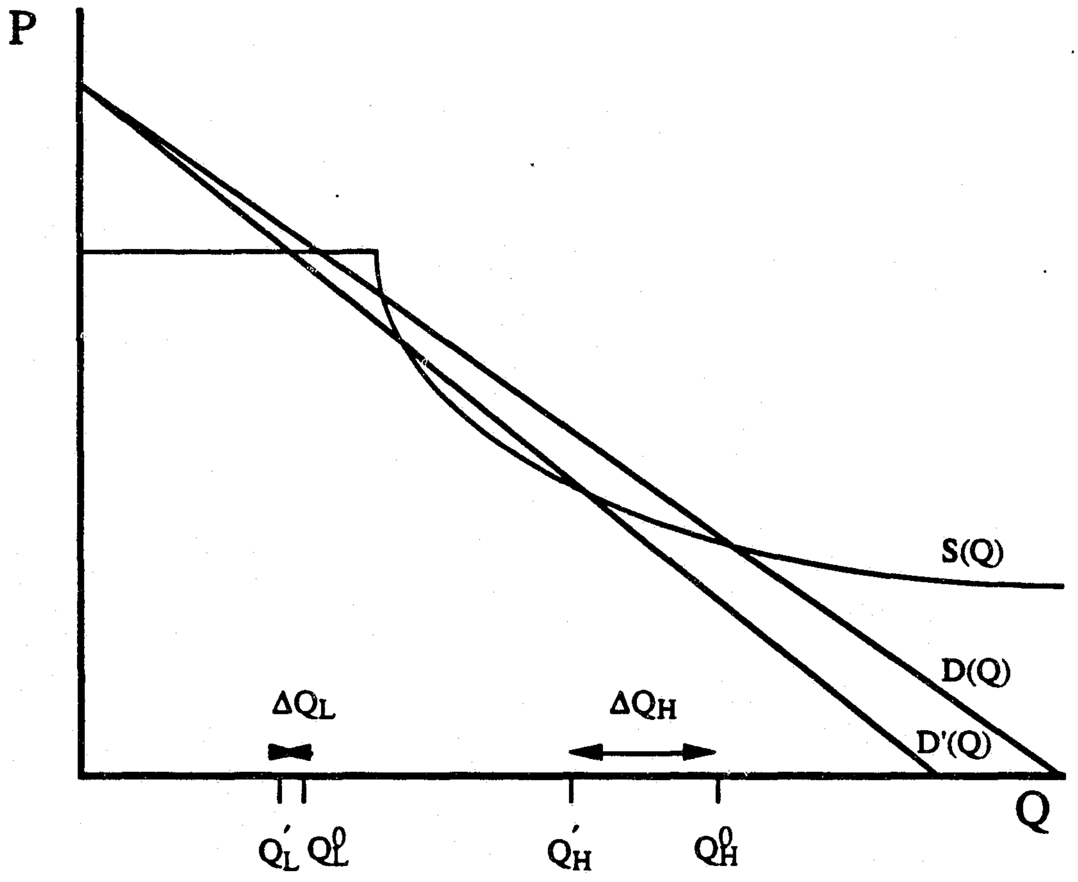
To summarize the predicted effect of harsh minimum sentences, if the criminal justice system's punishment capacity is already fully utilized, they will have no appreciable effect. If, on the other hand, the market is at point A, they might be able to reduce consumption as long as the criminal justice system was not near saturation. If the criminal justice system were near saturation, however, there is some danger that they might tip the market to a high volume equilibrium at point C.

7.2.5 The Effectiveness of Demand Reduction

Subsection 7.2.2 argued that increasing enforcement may be a relatively ineffective way to reduce consumption if the market is already at a high consumption equilibrium. This section suggests, in contrast, that demand reduction may be particularly effective precisely when the market is already at a high consumption equilibrium.

Figure 7.6 shows the effect of a 12.5% reduction in demand at all prices (moving demand from D(Q) to D'(Q)). The corresponding percentage reductions in the market equilibrium quantity are about the same at points A and C, so the absolute reduction in quantity is much greater from point C.

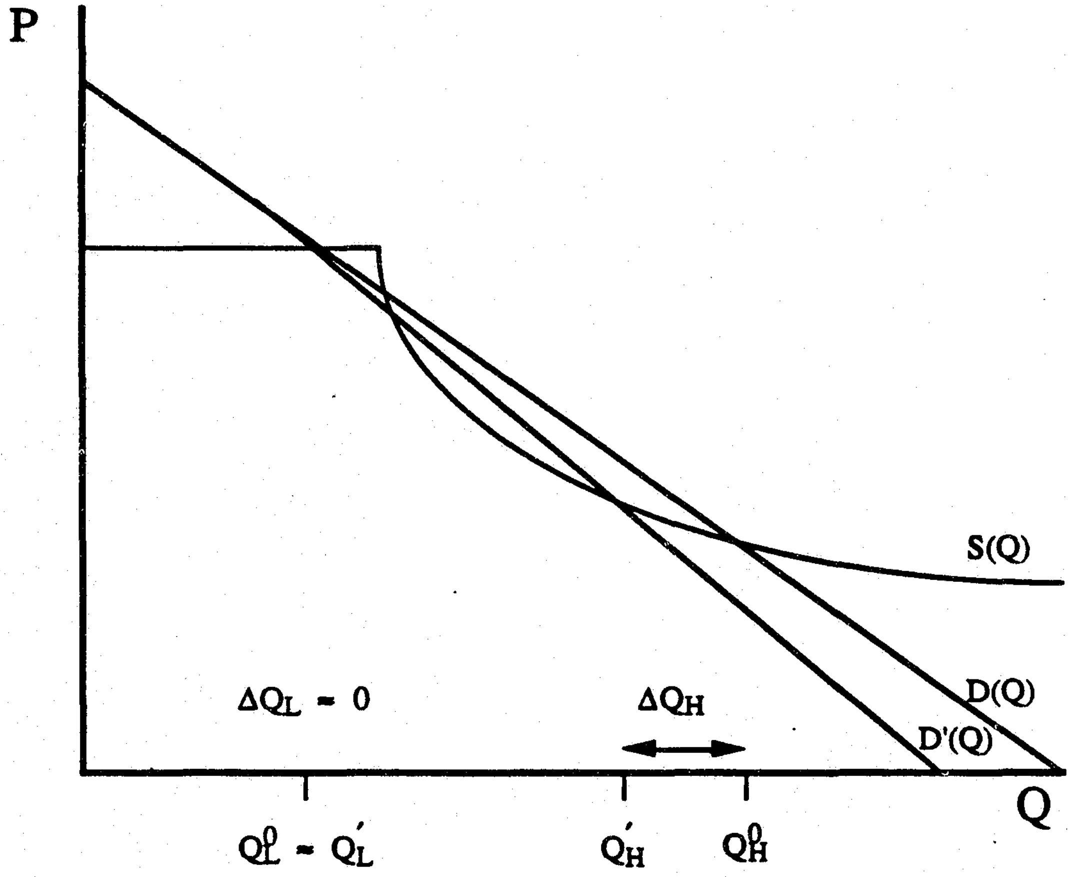
Figure 7.6:
The Effect of Reducing Demand By A Fixed Fraction At All Prices



Education about the health effects of illicit drugs may be more effective with "casual" users and potential "casual" users that with truly committed users. The committed users are the ones who contribute demand even at high prices. "Casual" users, as the term is used here, are those who consume if and only if the cost of doing so is not too great.

Then an education program might increase the downward slope of the demand curve beyond some point. This is depicted in Figure 7.7. It shows that if this were the case, the difference in education program's effectiveness at points A and C would be even more pronounced.

Figure 7.7:
The Effect of Reducing Demand of Users Who Are Not Addicted



Treatment programs for current addicts might behave in the opposite way. They might reduce the demand of committed users and hence appear as a vertical shift down in the demand curve. If that were the case, then treatment would be relatively more effective if the market were at the low quantity equilibrium, point A.

7.2.6 Summary

The multiple equilibria model has several policy implications. First of all, it suggests that if a community is at a low quantity equilibrium, it should respond quickly and decisively to changes in supply and demand that threaten to tip the market to a high quantity equilibrium like point C. The primary indicator that such a tipping is imminent would be the saturation or near saturation of the criminal justice system's punishment capacity. If one perceives that the national criminal justice system's punishment capacity is becoming saturated, one could reasonably infer that this is a pivotal time, and that expanding punishment capacity is vital.

If, on the other hand, one perceives that the criminal justice system has been saturated for some time and that the market has grown beyond the point at which it first became saturated, then one would suspect that the market has reached point C. In that case increasing enforcement would be relatively ineffectual. At point C enforcement makes a relatively minor contribution to the cost of using, and doubling something that is small accomplishes little.

If the market is at point C, the multiple equilibria model suggests stressing demand reduction instead. Demand reduction, especially in as much as it is relatively unlikely to sway committed users, is most effective at the high quantity equilibrium. Treatment, on the other hand, may have its greatest effect in reducing the demand by committed users.

Finally, the model suggests that minimum mandatory sentences are not a good idea. If the market were at point A and in no danger of saturating the criminal justice system, they might reduce consumption. But if the market is at point C they have no effect on consumption, and if the market is closer to point A but in danger of tipping to point C, then minimum mandatory sentences could push the system over the edge, leading to the much higher levels of consumption at point C.

7.3 Why Enforcement Is Not Futile

It has been suggested that if the market is at point C then enforcement is relatively ineffective at reducing consumption. This section points out reasons why it would be premature to conclude from this that it would be wise to end enforcement and legalize drugs.

7.3.1 The Market May Not Be At A High Volume Equilibrium

If one could say with certainty that the multiple equilibria model were accurate and the market were definitely at the high volume equilibrium, then one could say with some confidence that increasing enforcement would be ineffectual. However, it is not certain that either of these preconditions hold.

If the market were actually at point A then legalization could be disastrous. Removing enforcement in that case could move the market to the vastly higher level of consumption at point C. Then even if legalization were repealed and criminal sanctions restored, the market would not move back to point A. Legalization would be an irreversible experiment. Since the criminal justice system is not as overwhelmed as popular accounts suggest (See Chapter 2), it is possible that the market, or at least part of the market in some cities and towns, is not at point C.

Furthermore, the entire multiple equilibria model is speculative. It is reasonable, but no data of any kind have been presented to support it. Perhaps there are actually three stable equilibria and the market is currently at the intermediate equilibrium. Then legalization might push the market to the third, still higher level of consumption. Or perhaps the multiple equilibria model is simply wrong.

Whatever market equilibrium pertains, it is clear that the criminal justice system does impose some costs directly. Users and dealers, at least in some places, fear arrest. Removing that risk would lower costs and hence increase consumption. In Figure 7.1, the new equilibrium would be where the demand curve intersects the horizontal dashed line; this happens at a quantity greater than that corresponding to point C.

7.3.2 Drugs' Illegality May Constrain Consumption

Many of the costs that makes the supply curve as high as it is are attributable to dealer-dealer violence, robbery, fraud, and actions taken in response to these threats.[13] For the most part these costs would disappear if drugs were legal and dealers had recourse to the court system to make and enforce contracts. That is, even if no one were arrested on drug violations, the simple fact that drugs are illegal raises prices and hence reduces consumption.

The fact that drugs are illegal may also hold down demand by preventing advertising and stamping a mark of societai disapproval on the activity.

7.3.3 Enforcement's Indirect Effects on Costs

Illegality imposes three types of costs on drug use (1) the direct cost coming from arrest and punishment, (2) costs arising from the fact that drugs are illegal (discussed above), and (3) the indirect effects of arrest and punishment, which will be discussed next.

There are at least two indirect effects of enforcement. The first is related to the multiplicative model developed in Chapter 3. Suppose arrest and punishment raise the price at one level of the market. Raising the price there will raise costs further down the network because many of the costs of distributing drugs depend on

[13] Note, if a dealer defrauds or steals from another dealer, it is a transfer not a cost from the perspective of all dealers. However, many thefts are committed by non-dealers. Also, actions taken to prevent theft impose true costs, as does the additional uncertainty. And of course violence represents a true cost.

the drugs' value not just their quantity. So there is a multiplier effect. An enforcement cost of $1 at the import level will create several times that amount of additional costs in total.

The second indirect effect of enforcement is more subtle and more interesting. The greater the risk from enforcement, the longer and hence less efficient the domestic distribution network's distribution chains will be. Longer distribution chains lead to higher prices because the drugs are bought and sold more times before they reach the final user. Transactions are costly both directly in terms of time and effort and indirectly because they present opportunities for violence and fraud. The consequence of this is that long distribution networks increase the price required to call forth a particular quantity at the retail level.

To see this, note that middle-level wholesalers could bypass the retailers and sell directly to final customers. This would greatly increase the profit per transaction. They do not do this because it would increase their risk of arrest. Hence the branching factor of the distribution network, i.e., the number of people to whom a dealer sells, is determined by trading off profitability and risk.

To formalize this concept, think of a dealer who receives a given quantity of drugs and must decide to whom the drugs will be sold. For simplicity assume that the dealer will sell the same quantity to each of b customers. The price received clearly depends on b. If b = 1 then the customer would be at the same level as the first dealer, and so presumably would pay no more than the first dealer paid. If b = 10 say, then the dealer would be selling to lower-level dealers, and hence would receive a higher price per unit. If b were high enough that the dealer were selling to final customers, the dealer would receive the retail price.

Hence if $R(b)$ stands for the dealer's revenues as a function of the number of customers, $R'(b) > 0$.

It is probably also true that $R''(b) < 0$. Suppose that currently the branching factor were x at all levels and prices increased by $100\alpha\%$ at each level. Then

$$\begin{aligned} &R(x) = 1 + \alpha \\ &R(x^2) = (1 + \alpha)^2, \text{ and abstracting somewhat} \qquad (7.1) \\ &R(x^\beta) = (1 + \alpha)^\beta. \end{aligned}$$

So if $x > (1 + \alpha)$, which it almost certainly is, then $R''(b) < 0$.

The risk of arrest also rises with b. For simplicity model the risk of arrest for dealing with a customer as a Bernoulli random variable with probability p that the attempted sale leads to arrest. Further assume that these probabilities are equal and independent for all customers. Then if the dealer sells to b people, the probability of arrest is $1 - (1 - p)^b$.

Finally assume that the dealer has some utility function U(x) with the usual properties that it is increasing and concave ($U'(x) > 0$ and $U''(x) < 0$), and let $-C_F$ be the cost of being arrested.

Then to maximize expected utility, the dealer must solve the following problem

$$\max_{b \geq 0} \; [1 - (1 - p)^b]\, U(-C_F) + (1 - p)^b\, U(R(b)) \qquad (7.2)$$

The first order condition for this is

$$U\left(R\left(b^*\right)\right) - U(-C_F) = -\frac{U'\left(R\left(b^*\right)\right) R'\left(b^*\right)}{\ln(1-p)}. \qquad (7.3)$$

Consider how increasing the arrest risk p affects the optimal branching factor b*. As p increases, (1 – p) decreases, so ln(1 – p) becomes more negative. To maintain equality, either $[U'(R(b^*))\, R'(b^*)]$ must increase, or $[U(R(b^*)) - U(-C_F)]$ must decrease, or both. Since an increasing function of an increasing function is increasing, U(R(b)) is increasing, so decreasing b* decreases $[U(R(b^*)) - U(-C_F)]$. Likewise, a concave function of a concave function is concave, so decreasing b* increases $[U'(R(b^*))\, R'(b^*)]$. Hence b* decreases as p increases.

So the greater the arrest risk, the lower the optimal branching factor, and hence the longer the distribution chain. Note that if the current branching factor is obtained by some optimization (e.g. cost minimization), then by the envelope theorem this additional cost for small changes in enforcement would be negligible. For larger changes, however, it could be significant, particularly if the costs imposed by other participants in the market exceed the costs imposed directly by the authorities.[14]

To summarize, suppose the criminal justice system were to impose additional punishment for which high-level dealers would need to be compensated by one million dollars. A first order analysis would suggest that retail drug revenues would, ignoring changes in the quantity consumed, rise by about one million dollars to compensate the dealers for the extra risk they incur.

The analysis in Chapter 3 suggests instead that prices would rise by enough to generate x million dollars, where x is roughly the

[14] According to Garreau (1989), "Drug buyers and sellers believe they have more to fear from each other than from police."

ratio of retail price to the price at the level affected directly. This increase would compensate lower level dealers for the greater cost of distributing drugs that are worth more.

The argument here suggests that there would be yet another effect. The extra risk would induce the domestic distribution network to reduce its branching factor, increasing the average number of transactions required to deliver drugs to the retail level. Since transactions are costly (both in terms of time and in terms of increased risk and opportunity for violence) the total costs to the domestic distribution network would rise, making the distribution system less efficient and hence raising the retail supply curve.

7.4 The Demand For Illicit Drugs

7.4.1 The Effect of Addiction on the Demand for Illicit Drugs

As Sections 7.1 and 7.2 discussed, the supply curve for illicit drugs may not have the usual upward slope. This section argues that the demand curve may also have some unusual characteristics. In particular, the demand curve at any point in time may be a function of the quantity consumed in the past. The more consumption there was in the past, the more people are likely to be addicted today, and the more addicts there are, the higher the demand curve will be. This property helps explain the notion that the price elasticity of demand for illicit drugs is relatively small in the short run, but greater in the long run.[15]

One is normally reluctant to discuss changes in demand because changing demand can explain almost anything; it is difficult to devise hypotheses that can be contradicted. However, the demand for illicit drugs is clearly not constant. At least two causes of shifts in demand can be distinguished. For simplicity they will be referred to as fashion and addiction.

The fashion effect occurs because drugs' reputations change. For instance, cocaine was once seen as a drug for successful people; now it is more closely associated with violence and poverty. This change has probably affected demand for cocaine in middle class communities.[16] Likewise high school seniors' perceptions of the dangers of using marijuana have been growing and the prevalence of use has been declining.[17] One cannot be sure whether the first caused the second or whether the second reflects a change in

[15] This distinction is made by Reuter and Kleiman (1986, pp.298-300) and Reuter, Crawford, and Cave (1988, p.21-23) among others.

[16] Marshall, 1988b, p.1159.

[17] U.S. Department of Health and Human Services, 1988c.

demand, but those linkages are at least plausible. At a broader level, Musto[18] suggests that society as a whole goes through cycles of permissiveness and intolerance with regard to drug use. One aspect of those cycles is shifts in demand.

This section will not consider changing fashion because there does not seem to be much opportunity to address such changes quantitatively. Instead it will consider changes in demand resulting from addiction.

Addiction is difficult to define,[19] but a rigorous definition is not required here, just the notion that some users begin to value drugs more relative to other things (including money). Such individuals will demand more drugs at any given price or, equivalently, will be willing to pay more to obtain a given quantity of drugs. In other words, their individual demand curves shift out.

Note this is different than tolerance. The users of many drugs develop varying degrees of tolerance for those drugs and then require larger doses to achieve the same subjective effect.

When users develop tolerance, the benefit derived per unit of drug declines, so presumably their demand curve shifts back. Realistically, it is probably true that the demand curve for many people who are developing tolerance is shifting out not back, but that is not be a consequence of tolerance. Rather it indicates that the addiction effect is dominating the tolerance effect.

The aggregate demand curve is just the sum of the individual consumers' demand curves, so addiction can shift the aggregate demand curve out. Since the development of addiction is positively related to the consumption of drugs, this means that the demand curve for illicit drugs has the curious property of being a function of the quantity consumed in the past. The less that was consumed in the past, the less addiction there will be and hence the lower the demand curve will be.

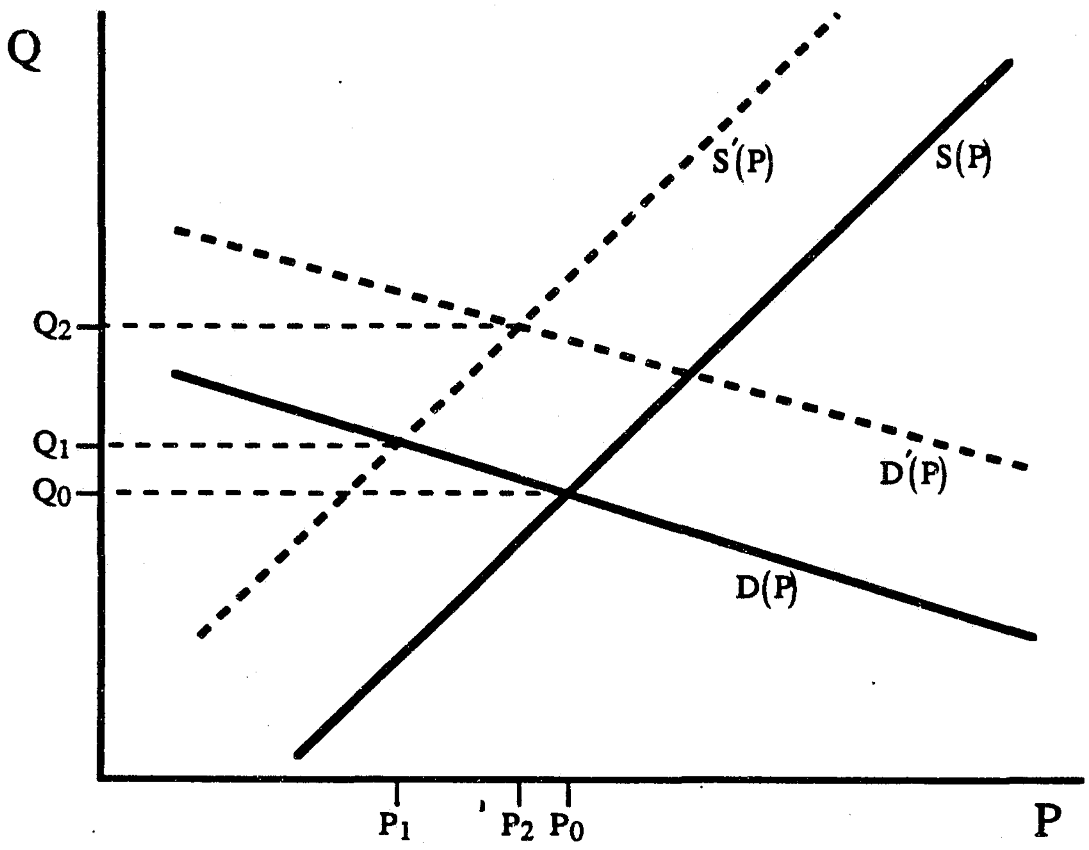
Broadly speaking this increases the apparent long run elasticity of demand. Suppose the supply increases, so more is offered at any given price. This is shown in Figure 7.8 as a shift from the solid upward sloping line (S(P)) to the dashed upward sloping line (S'(P)). (To emphasize the point that the unusual characteristics of demand discussed here are independent of the unusual characteristics of the supply curve discussed above, the supply curve is drawn with the usual upward slope.) When the supply curve shifts, prices fall and the quantity consumed increases from Q_0 to Q_1. But if increasing

[18] Musto, 1987.

[19] As is evidenced by the length of the addiction entry in The National Institute on Drug Abuse's (NIDA's) *Guide to Drug Abuse Research Terminology* (NIDA, 1982).

consumption leads to greater addiction, the demand curve will shift out too, moving from the solid downward sloping line to the dashed downward sloping line (from D(P) to D'(P)). This further increases the quantity consumed from Q_1 to Q_2, leading to more addiction and hence more demand and still more addiction. Presumably this positive feedback will die out; that is, each subsequent round of increases will be smaller than the previous one, but the net effect will be that when the supply curve shifts out, the quantity consumed increases more than one would anticipate looking only at the original demand curve.

Figure 7.8:
The Effect of A Shift in Supply on the Demand Curve



Furthermore, the equilibrium price will fall less than it would if demand were constant. It is even conceivable that demand could grow enough to push prices back up to their original level.

7.4.2 A Functional Form for the Demand Curve

This concept can be formalized with a model. There is no obvious way that the model can be tested, however, so it is best to think of it as a way of illustrating a point, rather than a way to describe the actual mechanics of the market.

Divide demand into two categories: demand from addicted users and demand from non-addicted users. Non-addicted users will be referred to as "controlled" users, but it should be understood that they include not only regular controlled users, but also occasional users, recreational users, and people who have not yet begun to use (some people who do not currently use contribute to demand at prices below the current price).

Suppose there are relatively few addicts, so the demand curve for controlled users is a function only of the price. (If there were many addicts that might reduce the size of the non-addict population and hence reduce their demand.) Suppose in particular that it has some constant elasticity β_c (the subscript c denotes "controlled" users), so the quantity demanded by controlled users as a function of the price P is:

$$Q_c(P) = \alpha_c P^{\beta_c}. \tag{7.4}$$

Suppose that addicts' demand also has a constant elasticity β_a (a for "addict"). Presumably addicts' demand is less price elastic than the demand from controlled users, but it is not perfectly inelastic, so

$$\beta_c < \beta_a < 0. \tag{7.5}$$

Now suppose that the quantity demanded by addicts is proportional to the quantity of drugs consumed in the past. One might at first think that it should be proportional to the quantity consumed by addicts, but some controlled users become addicts and addicts' consumption is probably not too different from total consumption. Even though there are more controlled users than addicted users, each addicted user consumes much more than a typical controlled user, so addicts' consumption probably dominates total consumption.

The way in which past consumption is measured will affect the model's behavior, particularly the rate at which demand adjusts to changes in consumption. For simplicity assume that time is discrete and Q_{t-1} is the amount consumed in the previous period. Then the amount demanded by addicts in the current period is

$$Q_a(P) = (Q_{t-1})(\alpha_a P^{\beta_a}). \tag{7.6}$$

And thus the overall demand curve is

$$Q_t(P) = Q_c(P) + Q_a(P) = \alpha_c P^{\beta_c} + (Q_{t-1})(\alpha_a P^{\beta_a}) \tag{7.7}$$

7.4.3 Short and Long Run Price Elasticities of Demand

This explicit expression allows one to compare the short run and long run price elasticities of demand. In the short run, the level of addiction is constant. This can be modelled by making Q_{t-1} constant. Then the elasticity ε is

$$\varepsilon = \frac{dQ_t(P)}{dP}\frac{P}{Q} = \frac{\alpha_c \beta_c P^{\beta_c} + Q_{t-1}\alpha_a \beta_a P^{\beta_a}}{\alpha_c P^{\beta_c} + Q_{t-1}\alpha_a P^{\beta_a}}$$

$$= \frac{Q_c \beta_c + Q_a \beta_a}{Q_c + Q_a}. \qquad (7.8)$$

Since β_c and β_a are the short run elasticities for the controlled and addicted users, the overall short run price elasticity of demand is just the weighted sum of the controlled and addicted users' short run elasticities, with weights equal to the fraction of consumption accounted for by the two groups. If addicts do in fact consume much more than controlled users, then the short term elasticity will be close to the short term elasticity for addicts, which is to say it will be rather small.

The long run price elasticity is likely to be higher. In the long run a new equilibrium would be reached in which $Q_{t-1} = Q_t(P)$. Hence

$$Q_t(P) = \alpha_c P^{\beta_c} + Q_t(P)\,(\alpha_a P^{\beta_a}) \qquad (7.9)$$

which implies that

$$Q(P) = \frac{\alpha_c P^{\beta_c}}{1 - \alpha_a P^{\beta_a}}. \qquad (7.10)$$

The subscript t has been dropped because it is an equilibrium quantity. This function is decreasing in P, as demand curves should be, and is convex. Taking the derivative with respect to P shows that

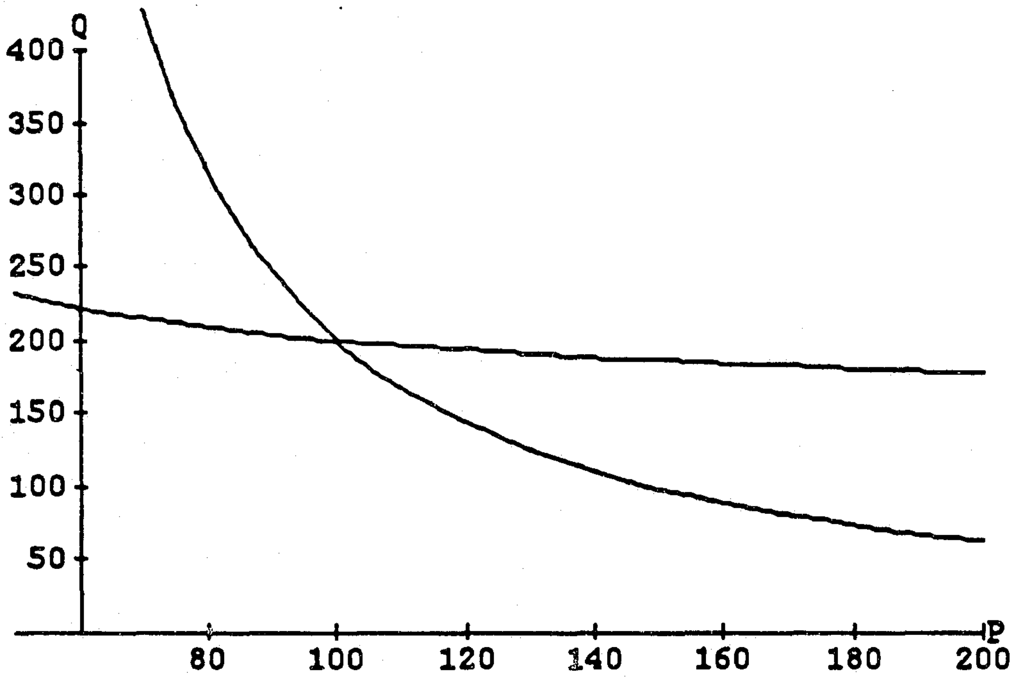
$$\varepsilon = \frac{dQ(P)}{dP}\frac{P}{Q} = \beta_c + \frac{f_a}{1 - f_a}\beta_a < \beta_c \qquad (7.11)$$

where $f_a = \alpha_a P^{\beta_a}$ is the fraction of consumption accounted for by addicts.

So, if demand is of the form described by Equation 7.7, the long run price elasticity of demand is even greater than the long run price elasticity of controlled users. If addicts consume most of the drugs (f_a is close to 1, say greater than 0.9), then the difference can be substantial.

Figure 7.9 shows the long and short run demand curves for particular parameter values. The values assume a short run price elasticity of −0.1 for addicts ($\beta_a = -0.1$) and −1.0 for non-addicted users ($\beta_c = -1.0$), and that at a price P of $100/gm total consumption is 200 tons (thinking of cocaine), and addicts account for 90% of consumption. Hence the vertical axis is measured in tons consumed, and the horizontal axis gives the retail price in dollars/gram. The much steeper slope of the long run demand curve is a reflection of its higher price elasticity.

Figure 7.9:
Long Run and Short Run Demand Curves



The discussion in the previous section showed that one supply curve can account for multiple stable market equilibria. Likewise, one expression for the demand curve (Equation 7.7) can explain low short run price elasticity and high long run elasticity.

7.4.4 Other Implications of the Demand Curve Equation

Equation 7.7 has some other interesting implications. For one, it suggests that at a given price the amount consumed by non-addicted users does not depend on addicts' demand parameters. For instance, making more treatment available to addicts might increase their elasticity of demand (increase β_a), but according to Equation 7.4 that would not affect the amount consumed by non-addicted users.

In contrast, the amount consumed by addicts in the long run

$$Q_a = \frac{\alpha_a P^{\beta_a}}{1 - \alpha_a P^{\beta_a}} \alpha_c P^{\beta_c} \qquad (7.12)$$

is affected by the demand parameters for non-addicted users. Reducing demand by non-addicted users by 10% will, in the long run, reduce demand by addicted users by 10%. In some sense this must be so because no one becomes an addict without first having been part of the demand created by non-addicts.

Hence, in the long run, the total quantity consumed will be proportional to the quantity consumed by non-addicted users. This suggests the importance of reducing the demand by non-addicted users.

Equation 7.7 also suggests that the long run equilibrium price will never be such that $\alpha_a P^{\beta_a} > 1$, i.e. the long run price will never be

$$P < \left(\frac{1}{\alpha_a}\right)^{1/\beta_a} = \alpha_a^{|1/\beta_a|}. \qquad (7.13)$$

Looking at Equation 7.7 shows why this must be so. If the price is low enough that $\alpha_a P^{\beta_a} > 1$, then the demand from addicts alone will exceed the total demand in the previous period. So, depending on the shape of the supply curve, the amount consumed and/or the price will rise, i.e. the previous price will not be sustained.

7.5 Summary

Sections 7.1 - 7.2 discussed the possibility that the supply curve is downward sloping. Section 7.4 suggested that the demand curve might shift over time when consumption changes, and that in the long run, the elasticity of demand might be fairly high. Taken together this suggests that the market for drugs may be highly unstable. Specifically, relatively small exogenous changes in supply or demand may lead to substantial changes in the quantity consumed, at least in the long run.

This is simultaneously encouraging and discouraging. It is encouraging because it suggests that well-planned government interventions may be able to accomplish something. It is discouraging because instability makes one wary of extrapolating past experiences to forecast the future; it adds uncertainty to a public policy issue that is already difficult to understand or manage.