THE DISTRIBUTION AND CONSUMPTION OF ILLICIT DRUGS: SOME MATHEMATICAL MODELS AND THEIR POLICY IMPLICATIONS

by

Jonathan P. Caulkins

B.S. Systems Science and Engineering, B.S. Computer Science,
B.S. Engineering & Policy, and M.S. Systems Science and Mathematics
Washington University in St. Louis, May 1987

S.M. Electrical Engineering and Computer Science
Massachusetts Institute of Technology, February 1989

Submitted to the Department of Electrical Engineering and Computer Science in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

Massachusetts Institute of Technology
May 4, 1990

Signature of Author_____
Department of Electrical Engineering and Computer Science
May 4, 1990

Certified by_____
Arnold I. Barnett
Professor of Operations Research
Thesis Supervisor

Accepted by_____
Amedeo R. Odoni, Co-Director
Operations Research Center

THE DISTRIBUTION AND CONSUMPTION OF ILLICIT DRUGS: SOME MATHEMATICAL MODELS AND THEIR POLICY IMPLICATIONS

by

Jonathan P. Caulkins

Submitted to the Department of
Electrical Engineering and Computer Science
on May 4, 1990 in partial fulfillment of the
requirements for the Degree of

Doctor of Philosophy in Operations Research

Abstract

The distribution and consumption of illicit drugs and the enforcement of drug laws have become serious societal problems. Nevertheless, relatively little is known about how the markets for illicit drugs operate or how government policies and other exogenous factors affect them. This thesis constructs mathematical models of various aspects of drug markets in an effort to at least partially remedy this deficiency.

The first chapter explains the approach taken; the second surveys existing sources of data on drug markets. The five subsequent chapters introduce models that address (1) how changes in import prices affect retail markets, (2) how so-called "open air" drug markets respond to local enforcement efforts, (3) how punishment policies influence the behavior of market participants, (4) how AIDS will affect the number of intravenous drug users, and (5) how the overall supply and demand for drugs differ from the supply and demand for more conventional products.

Thesis Supervisor: Dr. Arnold I. Barnett
Title: Professor of Operations Research

Acknowledgements

Thanks Mom and Dad for giving me the chance to write this thesis.

Thank you ~~Professor~~ ~~Barnett~~ Arnie for your advice and encouragement. Your praise and the generosity with which you gave your time constantly reinforced my belief in the value of this work.

Thank you Professor Kleiman for serving on my thesis committee, for sharing your experience, and for the many helpful suggestions which have been incorporated into this thesis.

Thank you Professor Larson for encouraging me to study this topic, for arranging my summer in Hartford, and especially for showing by example the value of writing an Operations Research dissertation on a subject that has not traditionally been explored quantitatively.

Thanks are due as well to:
Mark Moore for helping me get started in this field;
Peter Reuter for suggesting I study the subject matter of Chapter 3 and for responding to some of the ideas in Chapter 5;
Steve Morreale and John Coleman of the Drug Enforcement Administration's Boston Office for explaining how enforcement agents view the drug problem and for opening doors for me;
Maurice Rinfret for giving me data and information about prices;
Clifford Mullen for helping me obtain the rolling paper data;
Jack Homer for letting me use his data on cocaine prices and purities;
Ed Kaplan for his assistance with the AIDS models in Chapter 7;
Lieutenant Brian Kelly, Detectives Mark Lyons and Rich Perotta, and everyone else in the Hartford Police Department's Vice and Narcotics Division for letting me tag along with them so I could see some drug markets and local-level enforcement first hand.

Finally, thank you to all of my friends who helped me maintain (most) of my optimism and positive outlook on life even after spending so much time thinking about what is frequently a depressing and morbid subject.

This material is based upon work supported under a National Science Foundation Graduate Fellowship. Any opinions, findings, conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

Contents

Figures

Chapter 2

Chapter 7

Tables

Chapter 1: Introduction

1.1 Reasons for Studying This Topic

"Drug abuse" was the response given most frequently to an August, 1989 Gallup poll that asked Americans what they believed to be the most important problem facing the nation.[1] While polls can be misleading and public perceptions do not always mirror reality, few would dispute the assertion that drug abuse is a serious problem.

People are generally familiar with many of the problems caused directly by drug abuse. These problems include widespread addiction and dependence in the general population;[2] babies born addicted to drugs;[3] drug induced crime;[4] street violence leading to the death and injury of innocent bystanders,[5] police officers,[6] and participants in the drug trade;[7] overcrowded courts and prisons;[8] and the direct costs of operating the criminal justice system.[9]

There are less direct effects as well. These include political instability in source and transshipment countries,[10] high-level corruption in source and transshipment countries,[11] corruption in this country,[12] death threats against government officials in this country,[13] assassination and intimidation of officials in source countries,[14] environmental damage both in source countries[15] and in

---

[1]Isikoff, 1989b. Drug abuse retained this status at least until March, 1990 (Morin, 1990).

[2]U.S. Department of Health and Human Services, 1989c.

[3]Hundely (1989), Besharov (1989), and Kantrowitz et al. (1990).

[4]Gropper (1985) provides an introduction to this literature.

[5]Daley and Freitag, 1990.

[6]The brief biographies given with the list of International Narcotics Enforcement Offices Association 1989 Awards (The Narc Officer, November 1989, pp.34-97) make clear the risks to law enforcement officers.

[7]Reuter et al. (1990) note a street dealer in Washington D.C. is subject to a roughly 1.4% annual risk of homicide.

[8]Bureau of Justice Statistics (1989) and Pitt (1989).

[9]U.S. Department of Justice, 1989, p.3.

[10]This issue was a focus of a symposium entitled Drugs, International Security, and U.S. Public Policy held at Tufts University, February 27-March5, 1989.

[11]For an example of such charges see Branigin, 1989.

[12]Weld, 1988.

[13]Kouri, 1989.

[14]E. Robinson, 1989.

[15]Germani, 1988.

the U.S.,[16] diminished quality of life in neighborhoods around drug markets,[17] increased risks users pose to co-workers and the general public,[18] pressure to give up civil liberties (e.g. by relaxing rules of evidence),[19] and even occasional human sacrifice.[20]

The government has responded by vastly increasing budgets for anti-drug programs, particularly enforcement.[21] Many politicians have responded by calling for stiff sanctions against offenders, and the media has devoted considerable attention to drug issues.[22] In contrast, although researchers have studied the causes of drug abuse and drugs' pharmacological effects,[23] relatively few have studied the drug markets themselves or how various policies affect them. This has been noted by drug policy researchers,[24] drug policy makers,[25] and independent scholarly journals.[26]

The leaders of the small group that do study drug markets, Mark Moore and Mark A.R. Kleiman of the John F. Kennedy School of Government and Peter Reuter and his colleagues at the RAND Corporation, are economists and public policy analysts. This thesis attempts to complement and extend their work by taking an engineer's perspective. It uses the tools of Operations Research to develop mathematical models of the illicit drug industry and drug enforcement programs.

---

[16]See, for example Conners (1989). Also, The Department of Justice (1989) notes two examples in which animals were directly affected. In 1984 the use of poisons such as Warfarin and Havor by marijuana growers was responsible for the deaths of more than 1600 deer, and in Gelmer County, Georgia, a black bear died of an overdose after finding a duffel bag of cocaine in a clandestine landing area.

[17]Described by Kleiman (1988a) and cited as of extreme importance by Burke (1988).

[18]Schwartz, et al., 1989.

[19]Mydans (1989) and Morganthau et al. (1990).

[20]Woodbury (1989) describes a case in which some drug dealers kidnapped and sacrificed people as part of a religious ritual in the belief that it would bring protection from enforcement authorities.

[21]U.S. Department of Justice, 1989. This is discussed further in Chapter 2.

[22]Harwood (1989) describes the media's enthusiasm for covering drug issues.

[23]According to Barnes (1988b), "research on drug addiction is booming." Also, there are scholarly journals, such as the *International Journal of Addictions*, devoted to these topics, and they are the focus of much of the National Institute on Drug Abuse's (NIDA's) work. NIDA's narrow focus has stirred controversy, as is described by Booth (1988).

[24]Marshall (1988a) quotes Peter Reuter, Mark Moore, and James Stewart all making comments to this effect. See also Reuter (1984).

[25]The White House, 1989, p.87.

[26]Marshall, 1988a.

1.2 Why Studying Drug Markets Makes Sense

Just because an important problem exists and not enough people are studying it does not necessarily mean that Operations Research can contribute constructively to the policy debate. In particular, it does not assure that the problem is a suitable topic for a Ph.D. thesis in Operations Research. In this case, however, Operations Research can contribute on at least two levels: improving enforcement operations and improving the collective understanding of how drug markets operate.

Drug enforcement offers a wealth of interesting optimization problems. What, for example, are the optimal patrol patterns for Coast Guard cutters and border interdiction planes? Where should tethered aerostat balloons be deployed to maximize radar coverage? How should military resources be used to improve interdiction? Which vehicles and containers should receive special attention from Customs inspectors? How should information obtained by enforcement authorities be stored, disseminated, and managed? How can computer programs be designed to sort and sift data to find connections that provide investigative "leads"? All of these and other enforcement-management issues merit study, but they are not the focus of this thesis.

Instead, this thesis tries to explain some aspects of the behavior of drug markets. The reader may wonder about speaking of a "market" for a criminal activity, but this is one, although by no means the only, useful perspective on consensual crimes.[27] And the empirical evidence that is available supports the notion that one can speak of markets for illicit drugs.[28]

Drug use is a consensual crime because most illicit drugs are purchased in voluntary transactions. The distribution of drugs also has this character because dealers buy and sell from each other. Contrary to popular belief, drug distribution is generally not vertically integrated; there is no monolithic, tightly controlled entity that can set prices at will.[29]

---

[27]Economics has been applied to consensual crimes other than drug dealing. For example, Reuter (1983) and Reuter and Rubinstein (1979) used it to analyze the markets for bookmaking, numbers, and loan shark services.

[28]For example, Lisowski (1988, p.142) concludes that "analysis of wholesale cocaine price data supplied by the DEA supports the hypothesis that the data were generated by a market process in operation."

[29]See, for example, Reuter and Kleiman (1986), Reuter and Haaga (1989), and Reuter et al. (1990, p.29). An organizational unit one might think could exercise monopoly power is the Medellin Cartel, but it lacked the power to

Instead there are hundreds of thousands of individuals and small organizations in the business, at least on a part time basis, of producing and distributing illicit drugs. Many act primarily as brokers: buying, dividing, and repackaging drugs for resale at lower levels of the market. Others manufacture drugs. Still others are importers (smugglers). The majority, however, are retailers who sell directly to final customers. Essentially all of them are in business for profit, especially if one counts drugs withdrawn for personal consumption as part of their wages.

In many respects these drug producers and distributors behave like licit businesses.[30] They deal with customers, competitors, and suppliers; hold inventories; manage labor problems; and borrow and collect debts.[31] Likewise their customers in many ways behave like typical consumers.

Operations Research helps guide licit businesses operations, so its tools and philosophy translate fairly readily to illicit enterprises. The objective here is not to maximize the profits of these illicit enterprises, but rather to explain how they operate and to predict how they would respond to various stimuli, including changes in government policies.

Explaining how businesses react to stimuli is something microeconomists study. Indeed, any study of this kind will have a flavor of economics, and economic theory contributes to the understanding of drug markets. Illicit drug markets fall further short of the ideal market than the markets for most licit goods do, however, so describing them is not a solved problem.

For example, factors other than price affect behavior to a greater extent than is probably true of the markets for most licit goods. Principal among these are the risk of arrests, the health effects of using, and the risk of bodily injury or death resulting from the actions of other participants in the market. Also participants in drug transactions cannot enforce contracts in courts. This may sound like an arcane distinction, but it leads to widespread robbery, fraud,

---

prevent the precipitous drop in cocaine prices over the last ten years that almost certainly reduced its profits (Reuter, 1987).

[30]Garreau (1989) makes this point about dealers in his article "Washington's Underworld Entrepreneurs: Applying Free-Market Analysis to the Drug Trade in Our Nation's Capital." Preble and Casey (1969, pp.2-3) argue that even heroin addicts spend most of their time "aggressively pursuing a career that is exacting, challenging, adventurous, and rewarding," and that "taking care of business" is a term addicts frequently use to describe their activity.

[31]Compared with legal businesses, advertising and public relations are less important and distribution is more important, but those differences are relatively minor.

and misrepresentation of products' quality. A related issue is the imperfect flow of information. Clearly drug dealers cannot advertise, comparison shopping is difficult, and in general transaction costs are high. Finally, drug markets are characterized by a great deal of uncertainty. Obviously the threat of arrest is ever present, but there is even uncertainty about completed transactions. Restaurant patrons can reasonably expect a certain quality of food; not so drug customers. Drugs are frequently adulterated and sometimes the impurities are toxic. For all of these reasons there are aspects of the operation of drug markets that are not adequately explained by standard microeconomic theory.

Even if one concedes that drug markets are markets, one might still wonder about the wisdom of trying to describe people's behavior with mathematical models. After all, people's behavior is far more complex than that of even the most sophisticated factory robot, and there is little empirical data available to rein in excessive hypothesizing.

It is certainly doubtful that mathematical models of drug markets will ever yield quantitative precision approaching what is regularly achieved by models of physical systems. Nevertheless, one can identify at least six ways mathematical models can contribute to the policy debate surrounding illicit drugs.

The first is simply by providing a language for discussing the issues. The balloon model of local enforcement (Chapter 4) provides a rich vocabulary for describing some of the key aspects of local enforcement against so-called "open air" drug markets.

Mathematical models also encourage clear and careful thinking, including the frank discussion of assumptions. For example, the decision analytic framework Chapter 3 introduces helps identify the assumptions underlying two different models of how changes in import prices affect retail prices.

Mathematical models can also produce new insights. Afterwards it may be possible to understand and explain the insight verbally, but it is the modelling process itself that generates the insight. This is the case with the analysis of punishment policies that Chapter 5 describes.

Mathematical models can also help identify which parameters or phenomena are truly important and which play only a minor role. For example, the balloon model suggests that the elasticity of drug sales with respect to the number of dealers plays a key role in determining how a local drug market responds to enforcement pressure.

Models can also provide a visual image that conveys concisely a fairly complicated idea. This is true of both the multiple equilibrium model (Chapter 7) and the related concept of positive feedback (Chapter 4).

Finally, there are times when mathematical models can guide quantitative estimates. For example, Chapter 6 develops two models of the way AIDS will affect the number of intravenous (IV) drug users. They provide some basis for forecasting trends in AIDS case rates among IV drug users and the size of the IV drug using population.

Hence even in an area that has traditionally resisted quantitative analysis, mathematical modelling can make a useful contribution. Like all powerful tools, however, models can be misused. It is hoped that the text explains the models' limitations, but an overall caveat may be in order. Mathematical models are only models, and since models inevitably simplify, they can be misleading.

1.3 Approach Taken

It should be stated at the outset that this thesis will not present a single, unified model encompassing all aspects of "the drug problem." There are a variety of reasons for this, including the existence of multiple decision makers and multiple and sometimes conflicting objectives,[32] the inadequacy of the data,[33] and the relatively primitive understanding of the components of the "drug system" which makes it premature to attempt to model the system as a whole. The most fundamental reason, though, is that "the drug problem" is not a single, well-defined problem. Hence, there is no single "solution" in the same sense that an efficient algorithm can be said to solve a particular optimization problem. Rather, drug distribution, consumption, and enforcement constitute an application area that offers a variety of interesting problems just as other application areas, such as airline operations, do.

This thesis looks at a variety of issues related to illicit drugs, but of necessity it will not address them all. It neglects prevention and treatment because at least in some respects they seem to be less amenable to quantitative analysis. It avoids moral and ethical arguments and issues involving civil liberties because the objective is to inform the public policy debate, not to enter the debate on one

---

[32]See, for example, the discussion in Reuter, Haaga, Murphy, and Praskac,1988.
[33]This problem is discussed in Chapter 2.

side or the other. Finally, it does not discuss legalization because that subject has received considerable attention elsewhere,[34] and because it appears to be politically moot. Instead, it discusses aspects of the operation of drug markets. The next section describes what the thesis includes, but first the remainder of this section will explain a little more about the general approach taken.

The models developed are primarily descriptive not prescriptive. They try to describe how markets work. Often such descriptions lead directly to policy recommendations, but the models were not developed to solve particular policy problems.

The reason for this is simply that currently relatively little is known about how drug markets work, and some amount of "basic science" must precede the application of that science. This is perhaps most true of a field like physics; the United States is building the Superconducting Super Collider even though it is not clear how any knowledge derived from it will be applied. The situation is less extreme with drug issues because there is little mystery about what some of the applications are, but it is still true that basic understanding is a prerequisite for analyzing particular policies.

One problem with studying drug markets is the lack of data.[35] One chapter (Chapter 3) describes a widely accepted model, proposes an alternative, proposes an empirical test of the relative merits of the two models, and looks at historical data. Another chapter (Chapter 6) attempts to make a quantitative prediction about how AIDS will affect the number of intravenous drug users. Nevertheless, the majority of the models developed seek to describe the essential qualitative behavior, and do not rely on any specific source of data.

A reasonable response to the lack of data would be to keep the models very general, and in particular to avoid specifying functional forms. At times this approach is taken, but sometimes maintaining this generality leaves one unable to do more than show whether certain quantities are positive or negative. At those times I often propose what seems to be a plausible functional form and work with it, test its behavior against intuition, and analyze its implications for policy. This is a less cautious approach, but at times a more fruitful one. Frequently the behaviors modelled are rich, and simply

---

[34]The Data Center and Clearinghouse for Drugs and Crime even prints a bibliography of articles related to legalization and decriminalization, and an organization called the Drug Policy Foundation has been formed specifically to "reform" drug policy.

[35]Reuter and Kleiman (1986), Kleiman (1988b), and Reuter quoted in Marshall (1988a).

determining the sign of key quantities may fail to convey this complexity.

For example, the analysis in Chapter 5 uses an expression B(x) representing the benefit a typical user derives from using drugs as a function of x, the quantity of drugs consumed. Little is known about B(x). It is plausible that for a typical user it is an increasing function, and assuming there are diminishing returns to consuming larger and larger quantities, one could argue that B(x) should be concave. Beyond that, it is difficult to say much. Part of the analysis in Chapter 5, however, assumes that $B(x) = \sqrt{x}$ for $x > 0$, one function, but by no means the only one, which is both increasing and concave.

The intention is not to suggest that the relation modelled actually has the particular form proposed. Rather, the goal is to capture the essential characteristics of the true behavior while maintaining analytical tractability. Mathematical modellers always balance analytical tractability and realism. Since the existing literature on models of drug markets is slim at best, there are few guides as to what functional forms represent reasonable simplifications. Those selected here are consistent with available evidence and understanding, and the text discusses the likely effects of relaxing some of the assumptions, but it would clearly be valuable to generalize the models. If some of the assumptions are later shown to be unreasonable, then at least they may have helped stimulate that investigation.

An inspiration for this general approach comes from Jean Tirole, who revolutionized the field of Industrial Organization with his simple, stylized models.[36] His models provide a language for discussing key issues, develop intuition, and stimulate empirical research. It is hoped that one or more of the models presented in this thesis can achieve at least some of these objectives.

1.4 Overview of the Thesis

The next chapter describes what data and information are available about drug markets. It begins by discussing the problem of mythical numbers; the existence of numbers which are quoted authoritatively but are based on little if any evidence and may be seriously in error. It goes on to explain some reasons why gathering good data is difficult. The majority of Chapter 2 describes the data that are available on the criminal justice system (Section 2.3), production (Section 2.4), consumption (Section 2.5), and prices

[36]For a textbook-level treatment of his work, see Tirole, 1988.

(Section 2.6). Sections 2.7 and 2.8 briefly describe "case studies" of drug markets and some unconventional data sources, respectively. One unconventional data source, data on the production of rolling papers, has not, to the best of the author's knowledge, been used by any published studies.

Chapter 3 examines how changes in import prices affect retail prices. It has been argued that interdiction is futile because import prices are a small fraction of retail prices. Hence, even if interdiction doubled the import price, it would not appreciably increase retail prices. This argument assumes that price changes are passed along on a dollar for dollar basis, a view which is labelled the "additive" model. Chapter 3 suggests an alternate model, that price changes are passed along on a percentage basis. If this alternate view, called the "multiplicative" model, were true, then significantly increasing import prices would significantly increase retail prices.

Chapter 3 presents a decision analytic framework that suggests conditions under which these models might be expected to hold. Roughly speaking, the more expensive drugs are per unit weight, the more likely it is that prices will behave in ways similar to those predicted by the multiplicative model. For less expensive drugs such as marijuana or even cocaine before it has reached this country, one would expect prices to behave more like the predictions of the additive model. The limited data that are available seem to bear this out. Changes in retail cocaine prices were proportional to the corresponding changes in wholesale prices between 1982 and 1989, and the proportionality constant was more nearly equal to the ratio of the retail price to the wholesale price than to unity.

Chapter 4 turns to local drug markets. It introduces the balloon model, which describes how so-called "open air" drug markets respond to intensive local enforcement operations known as "crackdowns." The fundamental tenet of the model is that dealers can freely enter and exit markets. If the economic return from dealing in a particular market exceeds the return on dealers' next best activity, whether that activity is dealing in another market or not dealing at all, then more dealers will enter the market. If the return falls below the return available elsewhere, dealers will exit.

When enforcement against a particular market increases, the return to dealers in that market decreases, so some dealers exit. If their exit improves the lot of the remaining dealers then a new equilibrium may be attained. If their exit makes the remaining dealers worse off, still more dealers will exit, possibly creating a positive feedback effect that will collapse the market.

The model examines how the equilibrium market size varies with enforcement intensity, the conditions under which the market collapses, and the conditions under which a market might rebound. Some of the results obtained include an explanation of how gangs might play a role in the creation of new drug markets; some guidelines for choosing crackdown targets; and an explicit model of the positive feedback effect, which helps justify the very strategy of focusing resources on particular markets.

Chapter 5 explores how various punishment policies might affect the behavior of market participants, where a punishment policy specifies the expected cost of being arrested as a function of the quantity possessed at the time of arrest. The principal result concerns so-called controlled users who purchase some fixed quantity with a particular frequency. Intuitively one might expect that imposing the maximum possible punishment irrespective of the quantity possessed would minimize consumption, but that turns out not to always be the case. Reducing the punishment for smaller quantities may "bribe" users to purchase smaller quantities. Although the purchase frequency may increase, it need not increase enough to offset the decrease in the purchase size.

Under a set of plausible assumptions, the consumption minimizing policy is derived. Since it has some potentially objectionable characteristics, the consumption minimizing policy within a restricted class of policies that are more likely to be politically feasible is also derived. Implementation issues and simplifications inherent in the model prevent it from being used to quantitatively compute an improved policy, but it makes the point that in general it may be wise to make punishment an increasing function of the quantity possessed.

Chapter 6 develops two models that predict how AIDS will affect the number of intravenous (IV) drug users. The first is a simple population model; the second is an open population version of a model introduced by Kaplan.[37] Both concur that in the long run if other things (such as the price and availability of injectable drugs such as heroin) remain equal, AIDS could reduce the IV drug using population by 50% or more.

Chapter 7 takes a broad view of the supply and demand for illicit drugs. It argues that the supply curve may actually be downward sloping. That is, the cost per unit of supplying drugs may decrease as the market grows. A downward sloping supply curve can give rise to multiple stable market equilibria and hence could

---

[37]Kaplan, 1989.

explain both the high consumption at relatively low prices observed today as compared with 20 years ago without assuming gross shifts in supply and/or demand. The multiple equilibria model suggests that if the country is indeed at a high-volume, low-price equilibrium, then increasing enforcement is unlikely to significantly reduce consumption. In contrast, demand reduction efforts may be particularly effective at reducing consumption from such a point.

Chapter 7 next proposes a single functional form for the demand curve that yields a low short-run and high long-run price elasticity of demand. Several drug market researchers have predicted that there would be such a distinction between short- and long-run elasticities. The proposed functional form accounts for this behavior and offers an explicit explanation of why it might be so. Both of the models in Chapter 7 suggest that the market for illicit drugs may be unstable; relatively small exogenous changes could lead to large changes in the quantity consumed.

Chapter 8 summarizes the principal findings, attempts to extract some general conclusions, and offers some suggestions for further work.

The chapters in this thesis can be read in any order. Readers should feel free to skip sections in which the mathematics becomes to dense to follow; many of the key points can be grasped without mastering the derivations. Familiarity with the mechanics of the models may, however, help readers understand both their limitations and their potential extensions.

Chapter 2: The Data

2.0 Introduction

This chapter reviews the existing sources of data on drug markets. It has three goals: (1) to introduce the subject matter and give the reader a feel for the order of magnitude of various quantities, (2) to underscore the dangers of putting too much faith in some frequently quoted numbers, and (3) to describe what data are available.

The chapter describes some studies that are not data oriented if they represent "primary" research on the markets and their participants; it does not discuss studies that analyze or synthesize information gathered from other sources.

The next section describes the problem of "mythical numbers." All too often guesses and rough estimates are transformed into hard facts merely by force of repetition, especially when the numbers support a particular ideology or political position. So not only is there little hard data, but some of the numbers that are quoted authoritatively may be inaccurate.

Section 2.2 tries to explain why there is so little reliable data. Then the following sections describe some sources of information that do exist. Section 2.3 gives some statistics about the criminal justice system. Sections 2.4-2.6 describe information that is available about production, consumption and prices. The data on prices (Section 2.6) are particularly important because prices reveal a great deal about markets and because the following chapter develops a new theory for how import, wholesale, and retail prices are related. Section 2.7 gives examples of ethnographic and journalistic studies. Section 2.8 mentions a few unconventional and indirect data sources. One, sales of rolling papers, is a relatively unknown source.

2.1 Mythical Numbers

The policy debate concerning crime in general and drug problems in particular is fraught with "mythical numbers." Mythical numbers, such as estimates of the number of heroin addicts, usually reflect some underlying truth (the number of addicts is large and the estimates are large), but they are in some ways more like folklore than data. At best they are illustrative, but they certainly are not accurate.

This is a strong statement, and it may be slightly overstated, but unfortunately there is support for its basic tenor. Concern about mythical numbers shaped the very nature of this thesis. Since it is so central to my views, I will devote several pages to it, and I will use others' words to make the case.

In 1971 Max Singer wrote an article entitled "The Vitality of Mythical Numbers"[1] in which he showed that official estimates of the amount of property crime committed by heroin addicts had to be too high by at least an order of magnitude. He noted that:

> It is generally assumed that heroin addicts in New York City steal some two to five billion dollars worth of property a year, and commit approximately half of all the property crimes. Such estimates of addict crime are used by an organization like RAND, by a political figure like Howard Samuels, and even by the Attorney General of the United States.[2]

He then goes on to argue, however, that "if we credit addicts with *all* of the shoplifting, *all* of the theft from homes, and *all* of the theft from persons, total property stolen in a year in New York City amounts to some $330 million."[3]

Singer uses this calculation to argue that official estimates of the number of heroin addicts are too high and to point out that mythical numbers exists:

> This exercise is another reminder that even responsible officials, responsible newspapers, and responsible research groups pick up and pass on as gospel numbers that have no real basis in fact. We are reminded by this experience that because an estimate has been used widely by a variety of people who should know what they are talking about, one cannot assume that the estimate is even approximately correct.[4]

Back in the early 1970's public debate about drugs (although not necessarily the drug problem itself) was young, and people still had a lot to learn. One might expect that, even if mythical numbers existed 20 years ago, they would have long since been banished.

---

[1]Singer, 1971.
[2]Singer, 1971, p.3.
[3]Singer, 1971, p.5, emphasis in the original.
[4]Singer, 1971, p.6.

Peter Reuter argues in his 1984 article "The (Continued) Vitality of Mythical Numbers"[5] that this has not happened. Official estimates of the number of addicts, the amount of crime committed per addict, and the total amount of crime are still incompatible.

He notes that today, more sophisticated techniques are used to estimate the number of heroin addicts, but that sophistication does not always bring accuracy. Bruce Spencer independently arrives at the same conclusion in his paper "On the Accuracy of Estimates of Numbers of Intravenous Drug Users."

> Considerable ingenuity has been used by a number of researchers in attacking this difficult estimation problem. This paper does not say that the estimates should have been produced by alternative procedures; indeed, few constructive suggestions are made. Rather, it concludes that the estimates are simply highly inaccurate and form a weak basis for any policy or program decisions.[6]

Reuter concludes that: "At this stage the only respectable stance [with regard to the number of heroin addicts] is pure agnosticism."[7] He then goes on to criticize data on retail heroin prices, crimes committed by addicts, and the total dollar value of sales of marijuana, cocaine, and heroin.

How do these mythical numbers arise? It appears that sheer force of repetition converts rough estimates into conventional wisdom. Barnes, writing in *Science*, notes that at a National Institute on Drug Abuse (NIDA) meeting "health officials reported that there are 5000 new cocaine users daily, that 6 million people are regular users, and that 0.2 to 1 million are compulsive users."[8] She goes on to explain that:

> The notion that 5000 people each day try cocaine for the first time, for example, is based on telephone interview data first published in 1984. According to the best recollection of the writer who compiled these data for NIDA--and who currently works for the San Francisco-based "Just Say No" Foundation--the figures were extrapolated from the responses of 500 randomly selected callers who phoned a cocaine hotline in New

---

[5]Reuter, 1984.

[6]Spencer, 1989, pp.429-430.

[7]Reuter, 1984, p.142.

[8]Barnes, 1988a, p.1729.

Jersey. The estimated number of compulsive cocaine users is from the same publication.[9]

Needless to say 500 is not a large sample on which to base a number that received such widespread attention.[10] More importantly, callers to a cocaine hotline generate anything but a random sample.

Singer suggests in a footnote that: "Mythical numbers may be more mythical and have more vitality in the area of crime than in most areas."[11] He supports this assertion with two examples.

> In the early 1950's the Kefauver Committee published a $20 billion estimate for the annual "take" of gambling in the United States. The figure actually was "picked from a hat." One staff member said: "We had no real idea of the money spent. The California Crime Commission said $12 billion. Virgil Petersen of Chicago said $30 billion. We picked $20 billion as the balance of the two.
>
> An unusual example of a mythical number that had vigorous life--the assertion that 28 Black Panthers had been murdered by police--is given a careful biography by Edward Jay Epstein in the February 13, 1971, *New Yorker*. (It turned out that there were 19 Panthers killed, ten of them by the police, and eight of these in situations where it seems likely that the Panthers took the initiative.)[12]
>
> Reuter too gives an anecdote about a mythical number.
>
> My own favorite mythical number is the figure for the number of compulsive gamblers, a number that circulated in several official documents during the late 1960s and early 1970s. The figure, ten million, seems to have been based on a late-night phone call to a Gamblers Anonymous hotline, made by a

---

[9]Barnes, 1988a, p.1729.

[10]Nor is this the only instance in which national estimates have been extrapolated from small samples. Martz (1990) reports that the NIDA estimate of 862,000 weekly cocaine users (based on the National Household Survey) was projected from just 44 users in 8,814 households.

[11]Singer, 1971, p.6.

[12]Singer, 1971, p.6.

desperate government official who needed a number to fill out a table on the social costs of various behavioral disorders.[13]

Why do mythical numbers persist? Reuter suggests three reasons.

> First, there is no constituency for keeping the numbers accurate, while there is a large constituency for keeping them high. ... The second factor ... is the lack of any systematic scholarly interest in the whole issue. The literature on drug dealing, as opposed to drug use and the relationship between drug use and crime, is extremely slender. ... The third factor is most fundamental. The numbers have almost no policy consequence. It is certainly hard to identify any policy measure that rests on the estimate that the marijuana market generates $20 billion rather than $7 billion.[14]

Operations Researchers are used to taking data "with a grain of salt." Robust results are always valued and sensitivity analysis is important, but when studying illicit drug markets the usual scientific skepticism may need to be amplified; a little paranoia about the numbers is probably healthy.

Nevertheless this section is not meant to argue that the data can or should be ignored. Rather it is a plea to always know where the numbers come from, to take them as they are, and to resist the temptation to attach numbers to quantities that are simply not known.

If the National Household Survey reports that 1.0% of those surveyed reported that they had used heroin at some point in their lives, it means that when members of about 8,000 household in the continental United States were asked whether they had ever used heroin, about 1.0% said yes. It may mean that if one asked all of the members of all of the households in the continental United States, roughly 1.0% would say they had used heroin; whether that would be the case or not depends on how representative the sample of 8,000 households was and how much prevalences have changed since the survey was taken. It does not mean that 1.0% of the people in those households have used heroin because there may be under or over-reporting. And it certainly does not mean that one can estimate the number of heroin users by multiplying the population of the United

---

[13]Reuter, 1984, p.145.

[14]Reuter, 1984, pp.145-146.

States by 0.01; subpopulations excluded from the survey, such as the homeless, may have substantially different consumption patterns.

At times I use numbers in this thesis, including an estimate of the number of intravenous drug users, but I believe, and hope the reader also believes, that those numbers are highly suspect.

2.2 Why It Is Hard to Obtain Data

In economics there is a principle known as Gresham's Law which says that "Bad money drives out good." One hopes that the opposite occurs in science, that "good data drives out bad." The previous section argued that bad data circulates through the policy debate about drug markets. One may ask why this is so, or, more precisely, why there is not enough good data to drive out the bad. This section tries to give at least a partial answer to that question.

The problem is not simply one of incompetence or malicious efforts to hide the truth. The data are inadequate in no small part because it is difficult to measure most of the quantities of interest. This point is made in two ways. Subsection 2.2.1 examines one particular case, measuring the level of activity in an open-air drug market. It concludes that taking this measurement is not as easy as it might seem at first. Then Subsection 2.2.2 tries to extract from that discussion five broad reasons why studying illicit drug markets is difficult.

2.2.1 Example: Measuring the Activity in Open-Air Markets

Think for a moment about how one might go about measuring the level of activity in a single open-air drug market. As Chapter 4 will explain, there is a real need for such measurements. Local-level enforcement has been receiving considerable attention lately, but there is some debate about its effectiveness. Measuring the levels of activity in markets before, during, and after intensive local enforcement efforts could help administrators and researchers determine how effective those efforts are.

One obvious measure is the number of drug-related arrests made in the market over a given period of time.[15] It is equally obvious that this is a particularly poor measure. If the police crack down on a market, the activity there would almost certainly decrease, at least temporarily. One certainly would not expect dealing to increase; nevertheless, the number of arrests probably would.

---

[15]Morrison, Putt, and Zmud (1975) discuss this approach.

The reader might think that no one would try to estimate the level of activity from the number of arrests. Yet some estimates of the total volume of drugs consumed, and hence the total dollar value of sales, are based largely on seizure data. The parallel is strong. Seizures are not determined solely by the volume of drugs shipped. They also depend on the level of enforcement, the effectiveness with which enforcement resources are deployed, and dealers' and smugglers' skills at evasion. None of these factors are likely to remain constant, and the last two are themselves difficult to quantify.

Returning to the problem of measuring the level of activity in a particular drug market, one might consider monitoring citizens' complaints. For several reasons this is also at best an imperfect measure.

People might be more likely to notify the police if they think it will lead to some response, so there might be more calls during a police crackdown or public awareness campaign than there would be at other times.

Another problem with relying on neighbors is that they are not trained. They might overlook signs an experienced police officer would notice. Lack of training may also increase variability in observational ability. If dedication and observational powers vary from person to person, inter-market comparisons would not be possible. Finally, lack of training may introduce bias if neighbors are more likely to notice dealers who "look like dealers" or who are aggressive. Hence, they might perceive that a few highly visible dealers represent the same level of activity as many, less blatant offenders.

This raises an essential question. What is meant by the level of activity? Some possibilities include the number of dealers, the number of customers, the number of sales, and the dollar value of sales. These need not all be proportional to each other, so using different measures could lead to different results.

Furthermore, these measures are still vague. What exactly is meant by the number of dealers? The average number of dealers who make at least one sale on a given day or the average total number of active dealer-hours per day. Also, who gets counted as a dealer? Only the people who actually sell drugs, or should the count include people who direct customers to dealers (sometimes called "steerers"), look-outs, and "holders" (people, sometimes girlfriends, who actually hold the drugs so the dealer is less likely to be arrested for possession). One can ask similar questions about the other

seemingly precise measures. For example, are sales for different quantities and different drugs counted equally?

Problems with definitions are not confined to the measurement of open-air drug markets.[16] What, after all, makes someone an "addict"? Likewise, what transactions should be considered to occur at the "wholesale level"?

Returning to the problem of measuring activity levels, if there are problems with both police records and citizens' reports, one might consider instead using some relatively objective surrogate measures. For example, one could try to monitor changes in heroin consumption by counting the number of people nodding (sleeping off their high) in public. This measure has at least three serious flaws.

First, weather probably has as much to do with the number of people nodding in public as the quantity of heroin consumed. Second, it is not clear what fraction of heroin users ever nod on the street. Some regular users maintain an outwardly "normal" lifestyle. Others may be hard-core addicts that nod, but do so indoors. Others may never nod simply because that is not how they react physiologically to the drug. Third, unless all streets were surveyed, the measure would be vulnerable to changes in addicts' preferences about nodding places.

Another objective measure might be counting how many of the miniature zip-lock bags used to package drugs have been discarded on the street. Assuming there is no change in the fraction of zip-lock bags that are discarded as litter, this might give at least trend information about dealing. However, this measure also has problems.

Again weather could distort measurements. Bags could be hard to count on snowy days, and a stiff wind might scatter them. It may be that in bad weather more people go inside before taking their drugs, so the number of bags discarded might decline even if sales remained constant.

Also, the number and location of discarded zip-lock bags may have more to do with consumption in that area than with dealing. Finally, the dealers would almost certainly find out the police were counting discarded bags, and they could easily change their behavior either to intentionally confound measurements or because they mistakenly believed the discarded bags would be used against them as evidence.

Perhaps the most promising approach would be to simply watch the street from some concealed point. Even this would not be

---

[16]Singer (1971), Barnes (1988a), and Spencer (1989) all comment on the problems caused by imprecise definitions.

easy because accessible, concealed vantage points may not be available. It would, however, have the advantage of generating more detailed information than just the total level of activity.

The basic reason measuring the activity in a drug market is difficult can be explained by imagining that the events of interest occur according to a renewal process. If every event (arrival) were recorded, estimating the average arrival rate would be straightforward. But not all events are recorded. Let p(t) be the probability that an event would be recorded if one occurred at time t. If p(t) were constant, then the observed renewal process would also be a renewal process, and its average arrival rate could be estimated. Then one would merely need to estimate the constant p to obtain an estimate of the actual rate at which events occur. Even if one could never measure p, as long as it remained constant, one could still detect changes in the average rate of the underlying process.

Unfortunately p(t) is not likely to be constant for many measures one would like to use. For example, the probability a transaction results in an arrest increases during a crackdown. Likewise, the probability that neighbors report some activity may depend on their estimate of the probability that the police will respond.

The preceding discussion is intended to persuade the reader that measuring the level of activity in a single open-air market is difficult. Unfortunately most other quantities of interest are probably even harder to measure. Open-air markets are the easiest to observe. Measuring characteristics of so-called "quiet" dealing that takes place in bars, homes, and work places may be even more difficult. Likewise, the retail level is the most accessible to police. It is generally harder to gather information about wholesale dealers, smugglers, and "kingpins".[17] Furthermore, it would be helpful to know much more than the quantity transacted and the price. One would like, for example, to know a dealers' annual revenues and operating costs. That would reveal more about how enforcement affects dealers' actions than simply knowing quantities and prices.

The next subsection tries to extract from this discussion general observations about the difficulties inherent in studying illicit drug markets.

---

[17]A problem noted by Reuter and Haaga (1989) in their study of high-level dealers.

2.2.2 Reasons Why Data Collection is Difficult

The previous subsection discussed ways of measuring one quantity associated with drug markets. It noted that for a variety of reasons, taking this measurement is difficult. This subsection generalizes some of the points made above and lists five broad reasons why it is difficult to collect data on illicit drug markets.

Reason #1: It is hard to obtain a random sample.

Of necessity studies focus on subpopulations. Frequently these subpopulations are not representative of the larger population (for example, arrestees are almost certainly not representative of all users), and truly random samples even of these subpopulations are rare.

Reason #2: The quantities of interest are ill-defined.

Imprecise definitions lead to incomparable and sometimes distorted data. Many characteristics of interest, such as drug use, exist on a continuum, not at one of a few discrete levels. There is no clear distinction, for example, between controlled and uncontrolled use.

Reason #3: People collecting and giving data do not have an incentive to be accurate.

The extent to which under and over-reporting occur is not known, but it could be significant. Some users may boast about the size of their habit; others may under-report for fear of punishment. Enforcement agents may have an incentive to inflate arrest and seizure data if their career advancement depends on those numbers, and Reuter[18] notes that many agencies have an incentive to inflate estimates of the size of the drug problem.

Reason #4: Reporting standards vary over time.

Many data collection efforts stress their value for monitoring trends, perhaps partly because they cannot be calibrated to determine actual levels. However, if diligence in reporting and collecting data and/or definitions change over time, intertemporal comparisons may be deceptive.

Reason #5: Collecting good data can be expensive.

---

[18]Reuter, 1984, p.145.

The budget for research on drug issues, particularly on the operation of drug markets, is small. Since collecting good data can be expensive, this limits the amount of good data that is available.

There are problems with every approach one might take to studying drug markets. However, imperfect information is better than no information (at least if it is not misused), so various studies have been undertaken. Some of these are described next.

2.3 Criminal Justice System Statistics

The criminal justice system is an obvious place to look for information about drug markets. Records are kept on the number of people arrested, their offenses, the number of drug seizures, the kind and quantity of drugs seized, and so on. Three factors limit the value of this information, however.

The first is simply that the data are not always accessible. The criminal justice system is comprised of thousands of agencies: police and sheriffs departments, courts, and correctional facilities in hundreds of local jurisdictions and at the federal level. Information is scattered throughout the system. Collecting and combining information from more than one agency is time consuming. Agencies such as the Bureau of Justice Statistics do some of this, but they do not publish system-wide statistics on all of the quantities of interest to people studying drug markets.

Second, much of the data is about quantities that are of only secondary importance. For example, it is helpful to know how many people were arrested for dealing drugs in a given year, but it would be more valuable to know how many dealers there were in total.

Third, the mere fact that the authorities found out about a particular transaction or dealer introduces a bias. For example, it is useful to know the average price a DEA agent paid for a kilogram of cocaine, but it may or may not be the same as the market price.[19]

Criminal justice statistics do, however, contain useful information. This section will not attempt to review all sources of information about the criminal justice system; that would be a forbidding task, but it displays some relevant statistics in two categories: inputs and outputs. Subsection 2.3.1 briefly describes what resources have been allocated to drug programs, and how those resources have changed over time. Subsection 2.3.2 describes what

---

[19]Reuter, Crawford, and Cave (1988, pp.24-25) discuss this point and conclude that in this particular case the bias is probably not too severe.

those resources have "produced" in terms of arrests, convictions, seizures, and so on. Subsection 2.3.3 describes some of the data bases maintained by the Drug Enforcement Administration.

2.3.1 Inputs

Federal spending on drug programs has increased throughout the 1980's, and the increases have been quite dramatic in the last few years. (See Figure 2.1.[20]) Spending by state and local agencies has also increased.

Figure 2.1
Federal Drug Control Program Spending Authority, 1981-90 in Millions of Dollars

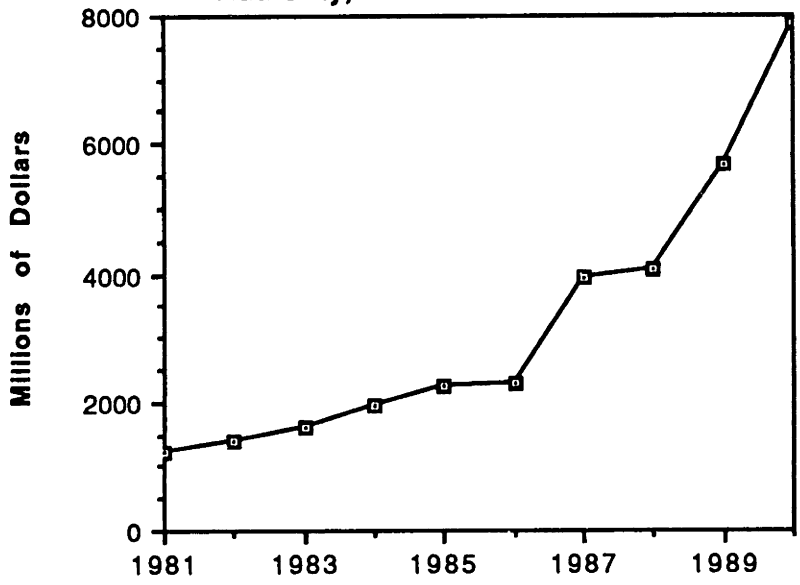


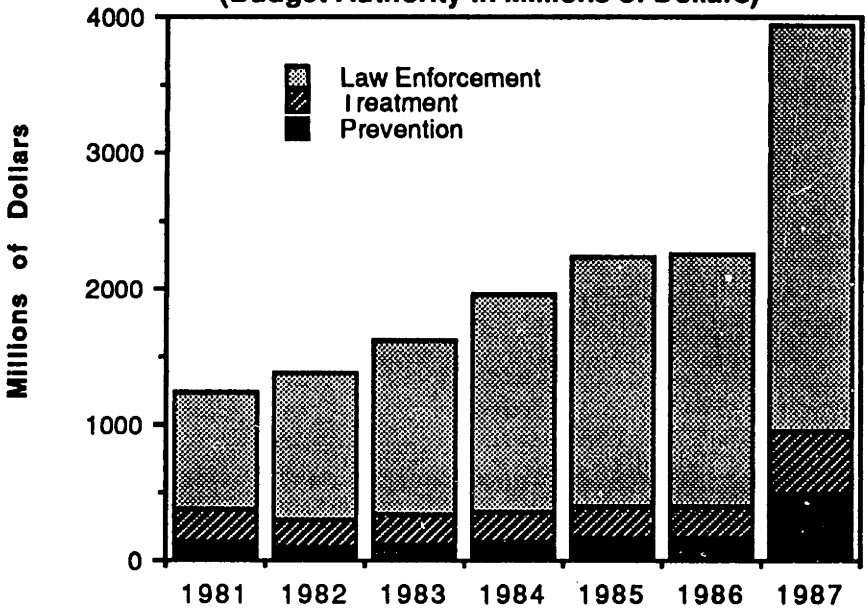
Figure 2.2[21] shows how this spending was divided between the broad categories of enforcement, treatment, and prevention.

---

[20]Source for 1981 - 1987 data: National Drug Enforcement Policy Board, *National and International Drug Law Enforcement Strategy* (1987, pp.182-183). Source for 1988 - 1990 data: The White House, 1989, pp.122-123. The FY 1990 number is the one proposed by William Bennett on September 5, 1989.

[21]Source for 1981 - 1987 data: National Drug Enforcement Policy Board, 1987, pp.182-183. The FY 1987 numbers are the amounts budgeted.

Figure 2.2
Division of Resources Among Federal Drug Control Programs, 1981-87
(Budget Authority in Millions of Dollars)



Federal enforcement spending grew throughout the 1980's, but spending on "demand-reduction" stagnated until 1987, when it too began to increase substantially. Nevertheless, enforcement still receives the majority of federal drug control spending.

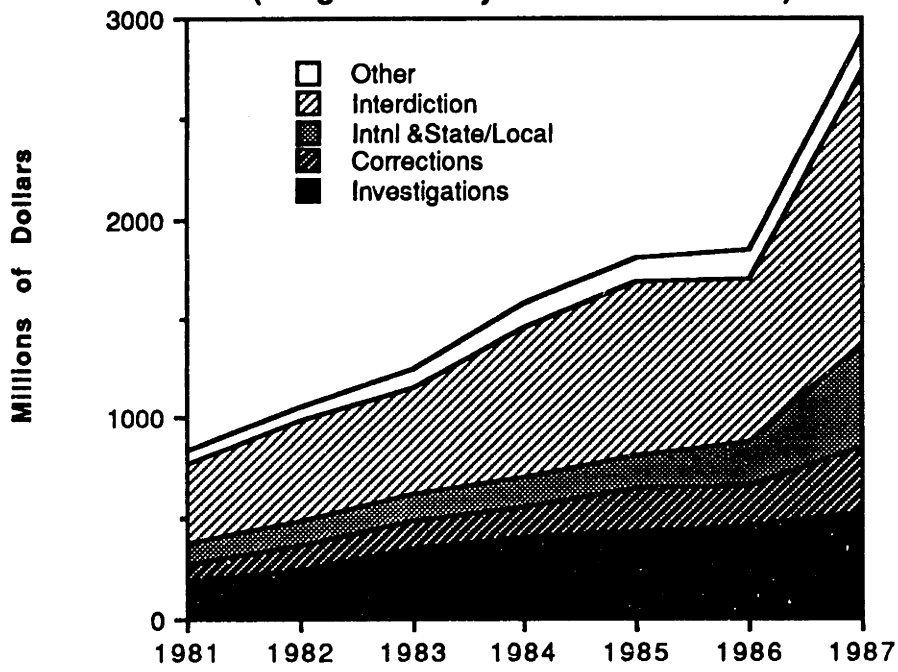
Figure 2.3[22] shows how these enforcement dollars are allocated. Interdiction receives the largest share, followed by investigations. Research and development spending remained at about 0.6% of enforcement spending throughout the 1980's, and never more than about a quarter of that went to Office of Justice Programs.[23]

---

[22]Source for data: National Drug Enforcement Policy Board, 1987, pp.185-187. The FY 1987 number is the amount budgeted.

[23]National Drug Enforcement Policy Board, 1987, pp.186-187.

Figure 2.3
Allocation of Federal Drug Enforcement Resources, 1981-1987
(Budget Authority in Millions of Dollars)



Intnl.&State/Local = International plus State and Local Assistance
Other includes prosecution, intelligence, and R&D

Thus spending on drug programs, particularly enforcement (and especially interdiction) increased throughout the 1980's. The next subsection looks at what this spending produced.

2.3.2 Outputs

Arrests and convictions are two of the principal products of the criminal justice system. Figure 2.4[24] shows the number of adults arrested by state, local, and federal agencies for drug violations between 1980 and 1987. Figure 2.5[25] shows the number of federal defendants convicted of drug offenses. Three things are obvious from the figures.

First, state and local agencies make more arrests than the federal agencies.[26] Second, most state and local arrests are for possession, while most federal convictions are for distributing and

---

[24]Bureau of Justice Statistics, 1989, Table 13, p.7.

[25]Bureau of Justice Statistics, 1988, Table 5, p.4.

[26]This can be inferred from the data because most federal drug arrests result in convictions.

manufacturing drugs. Third and most important, arrests for all kinds of drug offenses have been increasing steadily and substantially.

Figure 2.4
Adult Arrests for Drug Violations, 1980-87

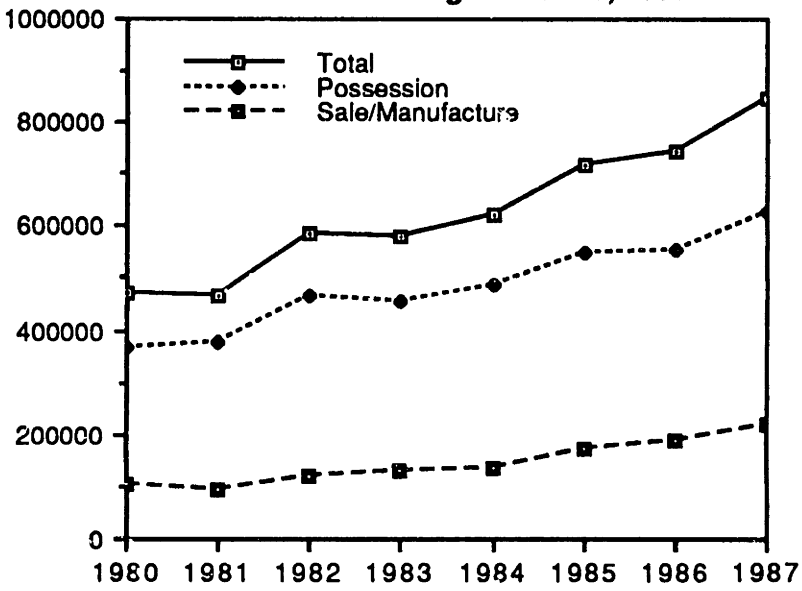


Figure 2.5
Number of Federal Defendants Convicted for Drug Offenses, 1980-86

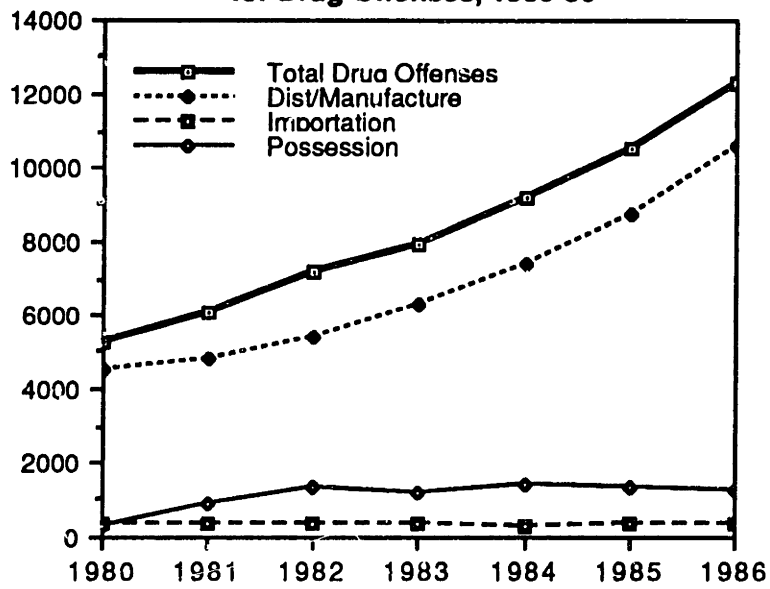
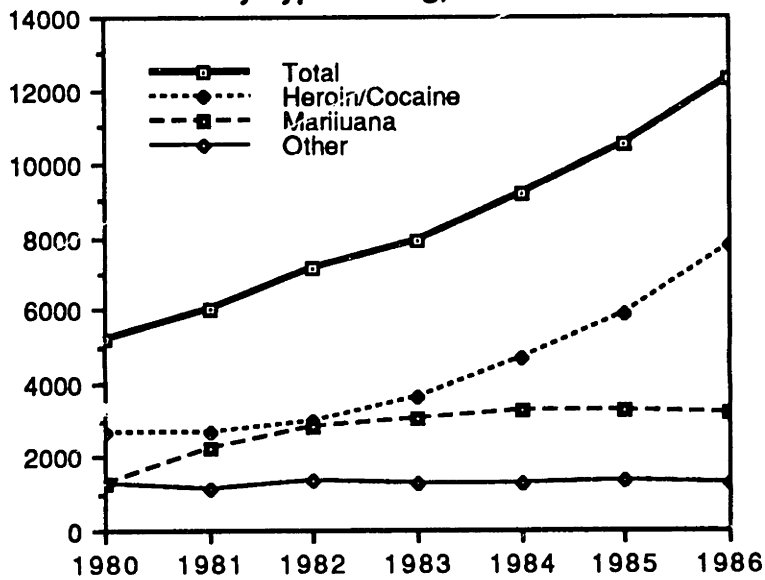


Figure 2.6[27] shows how these increases for federal enforcement were distributed among various types of drugs. Between 1980 and 1982 much of the increase was in marijuana convictions, but from 1983-1986, the bulk of the increase came from convictions for cocaine and heroin.[28] Note that there are many more federal drug law convictions for cocaine and heroin than for marijuana even though there are probably more marijuana dealers.[29]

Figure 2.6
Convictions of Persons Charged with Federal Drug Law Violations by Type of Drug, 1980-86



Arrests and convictions are not the only products of the criminal justice system's anti-drug efforts. Drug seizures are also monitored. Table 2.1 shows how much heroin, cocaine, and marijuana were seized through interdiction. Customs, Coast Guard, and INS numbers cannot be added because more than one agency may claim credit for joint operations. It appears, however, that seizures of heroin remained fairly stable and marijuana seizures probably declined, but seizures of cocaine grew enormously.

---

[27]Bureau of Justice Statistics, 1988, Table 6, p.4.

[28]Data from the National Drug Enforcement Policy Board (*Progress Report 1987*, July, 1988, p.75) on the number of people imprisoned as a result of DEA actions suggest that most of the increase was in fact for cocaine charges.

[29]Reuter and Kleiman (1986, p.294) estimate the number of marijuana, cocaine, and heroin dealers.

Table 2.1:
Illegal Drugs Seized Through Interdiction*[30]
(weight in pounds)

	Heroin	Cocaine	Marijuana
U.S. Customs**			
1983	594	19,602	2,732,974
1984	664	27,526	3,274,927
1985	784	50,506	2,389,704
1986	692	52,521	2,211,068
1987	639	87,898	1,701,150
U.S. Coast Guard			
1983		55	2,299,825
1984		1,932	2,857,511
1985		5,890	1,952,076
1986		7,495	1,840,678
1987		12,930	1,298,095
Immigration and Naturalization Service			
1983	11	154	38,700
1984	27	236	37,342
1985	23	1,378	72,473
1986	62	2,763	143,339
1987	83	13,121	226,055

*Due to differences in accounting methods, numbers in common categories cannot be added to arrive at an aggregate for all Federal agencies.
**These data include all seizures by Customs alone and, in many instances, in conjunction with or by other agencies.

Very roughly the same pattern holds for domestic seizures of these three drugs by the DEA and FBI. Domestic seizures of other kinds of drugs varied considerably from year to year.[31]

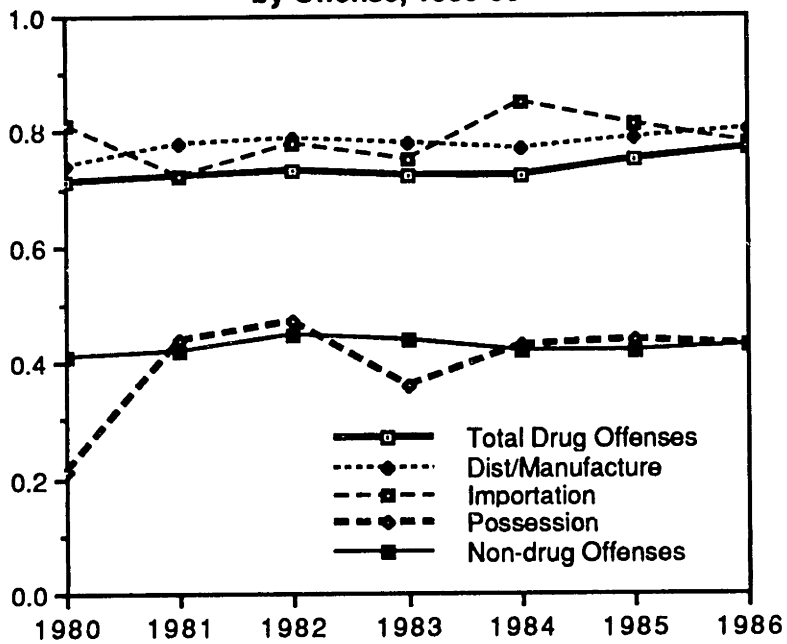
It has been widely suggested that the dramatic increase in arrests and convictions depicted in Figures 2.4-2.6 has swamped the criminal justice system. Some say that there is no room left in prisons, so that even if people are convicted for drug violations, they are not punished. Indeed federal, state, and local correctional facilities are filled to capacity and beyond.[32] But Figure 2.7[33] shows that at least at the federal level, the chance that someone convicted of a drug offense will be incarcerated has remained constant throughout the 1980's, is quite high, and is in fact considerably

[30]Taken from National Drug Enforcement Policy Board, 1988, p.39.
[31]National Drug Enforcement Policy Board, 1988, p.71.
[32]Bureau of Justice Statistics, 1989, Table 8, p.5.
[33]Bureau of Justice Statistics, 1988, Table 9, p.5.

higher than the corresponding chance for someone convicted of a non-drug offense.

This does not mean that drug-related arrests have not affected the criminal justice system. Even federal prisons are filled to capacity, and it is generally believed that most state and local systems are under even greater stress.

Figure 2.7
Percent of Convicted Federal Offenders Sentenced to Any Period of Incarceration by Offense, 1980-86



The increase in arrests and convictions might suggest that drug dealing has increased substantially in the 1980's, but such a conclusion depends on the assumption that more or less the same fraction of dealers are arrested each year. Another explanation for the increase is that the number of dealers has remained constant, but that enforcement efforts have increased, so a larger fraction of dealers have been arrested.

To summarize then, spending on drug programs increased substantially and so did conventional measures of the "product" of enforcement. It is not clear from the data, however, whether the size of the market grew or not. A priori, one would expect that, if all other factors remained equal, the market would shrink when enforcement efforts increased, but over a period of years, other factors, including demand and international supply, may well have changed.

2.3.3 The Drug Enforcement Administration's (DEA's) Data

The Drug Enforcement Administration maintains a number of data bases, at least three of which could be potentially useful to people studying drug markets.[34]

The Narcotics and Dangerous Drugs Information System (NADDIS) is the DEA's major enforcement support system. It has well over a million records on persons, businesses, ships, and airfields.

The Pathfinder II system includes information on individuals, events, aircraft, and vessels, including movement reports.

The System to Retrieve Information from Drug Evidence (STRIDE) stores information about the identity, quantity, purity, and (where relevant) the price of all drugs seized or purchased by the DEA. Data from STRIDE will be used in the next chapter.

Unfortunately much of the DEA's information is not available to the public.

2.4 Data on Production

The principal source of information about the production of illicit drugs is the National Narcotics Intelligence Consumers Committee (NNICC). NNICC was established in April, 1978 "to coordinate the collection, analysis, dissemination, and evaluation of strategic drug-related intelligence, both foreign and domestic, that is essential to effective policy development, resource deployment, and operational planning."[35] In 1988 its membership includes the Central Intelligence Agency, Coast Guard, Customs Service, Department of Defense, Drug Enforcement Administration (DEA), Federal Bureau of Investigation, Immigration and Naturalization Service, Internal Revenue Service, National Institute on Drug Abuse, Department of State, Department of the Treasury, and the White House Drug Abuse Policy Office.

NNICC publishes annual reports describing the trafficking situation. The 1988 report contained chapters devoted to cannabis, cocaine, dangerous drugs, opiates, and drug money. The chapters devoted to particular drugs and groups of drugs generally contained information from DAWN (described in Subsection 2.5.3); price

---

[34]The information in this subsection is taken from Drug Enforcement Administration, 1985.

[35]National Narcotics Intelligence Consumers Committee (NNICC), 1988, p.i. Information about NNICC reported here (unless otherwise noted) is from NNICC, 1988 and 1989.

information (presumably primarily from the DEA, whose price data are described in Section 2.6); and estimates of the quantities produced, seized, and imported.

Except for the DAWN data, the report does not describe how the estimates were made. Where ranges are given instead of point estimates, there is no indication of what confidence interval the range represents. In general there is no way to assess the validity of the estimates, but the recent upward revision by a factor of ten in estimates of Mexican marijuana production[36] suggests that they can differ considerably from the true number.[37] Presumably seizures data are reasonably accurate.

Even though information about quantities is uncertain at best, some of it is displayed in Tables 2.2 - 2.4 because the NNICC report is one of the only sources for this information.

Table 2.2 shows that roughly one-quarter of the marijuana consumed in the U.S. is produced domestically. Domestic producers account for an even larger fraction of THC (the principal psychoactive ingredient in marijuana) because domestic producers concentrate on high-potency sinsemilla varieties.[38]

Table 2.2:
Sources of Marijuana Available in the United States in 1987[39]

Country	Quantity (metric tons)	Percentage of Total Supply
Mexico	3,100-4,200	27.9
Colombia	2,300-6,600	32.5
Jamaica	145-285	1.7
Belize	200	1.6
Domestic	3,000-3,500	24.9
Southeast Asia	500-1,000	5.7
Other Latin America	500-1,000	5.7
Gross Marijuana Available:	9,545-16,585	
Less Seizures	3,000-4,000	
Net Marijuana Available	6,545-12,585	

[36]Sciolino, 1990.

[37]Reuter, Crawford, and Cave (1988, pp.73-77) also question some of the estimates of quantities imported.

[38]See Kleiman (1989) for a discussion of all aspects of the marijuana problem.

[39]Taken from the 1987 NNICC report Figure 3, pp.18-19.

Table 2.3:
Estimated Maximum Cocaine HCl Production by Country, 1987[40]

Country	Estimated Coca Leaf Yield (metric tons)	Maximum Cocaine HCl Capacity (metric tons)
Peru	98,000-121,000	196-242
Bolivia	46,200-67,200	92-134
Colombia	16,000-20,000	32-40
Ecuador	840	2
Maximum Cocaine HCl Production		322-418

The information about cocaine (Table 2.3) is probably well-known, but the information about heroin (Table 2.4) has some interesting implications. First, opium production is rising. Second, countries can quickly increase their production. In several instances production doubled from one year to the next. This suggests that even if production were halted in one country, production in another could be increased quickly to meet demand. Third, vastly more opium is produced than is needed to meet U.S. consumption. Assuming the U.S. consumes about 6 tons of heroin annually and it takes roughly ten tons of opium to produce one ton of heroin,[41] world production of opium is 30 to 50 times what is needed to meet U.S. demand.

Table 2.4:
Opium Production (metric tons)[42]

Country	1984	1985	1986	1987
Mexico	21	28.4	20-40	45-55
Burma	740	490	700-1,100	925-1,230
Thailand	45	35	20-25	20-45
Laos	30	100	100-290	150-300
Afghanistan	140-180	400-500	500-800	400-800
Iran	400-600	200-400	200-400	200-400
Pakistan	40-50	40-70	140-160	135-160
Total	1,416-1,666	1293.4-1623.4	1,680-2,815	1,875-2,990

To summarize, NNICC is perhaps the only official source of information about the quantities produced and imported, but the accuracy of its data has been questioned. Since it does not document its estimates, it is difficult to know how much faith to place in them.

---

[40]Taken from the 1987 NNICC report Figure 7, p.33.

[41]This conversion figure was used by Reuter and Kleiman (1986, p.293).

2.5 Data on Consumption

This section describes some data about the consumption of illicit drugs. Much of this information comes from surveys. Surveys face the problem of under- and over-reporting because respondents do not always have an incentive to answer accurately and memories are imperfect. Obtaining a random sample is also problematical. Some surveys try to cover whole segments of the population, but even those may omit important subpopulations. For example, the National Household Survey (described in Subsection 2.5.1) ignores the homeless and the High School Senior Survey (described in Subsection 2.5.2) misses drop-outs. Other surveys sample particular populations such as clients admitted to treatment centers (Subsection 2.5.4) or arrestees (Subsection 2.5.5). They may provide useful information about those subpopulations, but there is no way to generalize the results to the market as a whole.

2.5.1 The National Household Survey

The National Institute on Drug and Alcohol Abuse (NIDA) has, with assistance from other organizations, surveyed American Households about drug use nine times since 1971.[43] The survey covers Americans twelve years old and older living in the continental states. It excludes children under twelve, people living in institutions (military installations, dormitories, prisons, and so on), and transient populations (principally the homeless).

Excluding the homeless affects some estimates more than others. The homeless cannot represent a large fraction of marijuana smokers. Even if every homeless person smoked marijuana, their numbers would still be dwarfed by those included in the survey. On the other hand, excluding the homeless could significantly affect estimates of heroin use. It is quite possible that a significant number of heroin users are homeless. This compounds the problem that a Household Survey will produce more reliable estimates for drugs

---

[42]Taken from the 1987 NNICC report, Figures 16-18, pp.61-70.

[43]The first two studies (1971 and 1972) were conducted by the National Commission on Marijuana and Dangerous Drugs. The 1985 and 1988 surveys were supported by the National Institute on Alcohol Abuse and Alcoholism, and the 1988 survey received support from the Department of Education. Unless otherwise noted, all information in this subsection comes from U.S. Department of Health and Human Services, 1989c, *National Household Survey on Drug Abuse: Population Estimates 1988.* More detailed descriptions of the survey itself appear in the "Main Findings" documents such as U.S. Department of Health and Human Services, 1988b.

with higher prevalences, and the prevalence rates for marijuana are far greater than those for heroin.

Excluding college students and residents of Alaska and Hawaii may not affect national prevalence estimates severely, but it could reduce the survey's ability to forecast trends if those subpopulations are bellwether groups. This appears to have been the case with a smokeable form of amphetamines known as "ice"; it is fairly common in Hawaii, but is only beginning to reach the continental states.[44]

The 1988 data were based on 8,814 interviews with a 93.3% completion rate for households and a 74.3% response rate for individuals. The percentage and total number of people who have ever used, used in the past year, and used in the past month are reported for a variety of drugs in total and stratified by age, race/ethnicity, and region of the country. The drugs reported in this format include illicit drugs, (any illicit drug, marijuana, cocaine -- including crack, crack separately, inhalants, and hallucinogens -- including PCP), the nonmedical use of psychotherapeutic drugs, (any drug, stimulants, sedatives, tranquilizers, and analgesics), and licit drugs (alcohol, cigarettes, and smokeless tobacco). Both percentages and population estimates are given. The estimates are accompanied by 95% confidence intervals based on a logit transformation. Abbreviated tables are also given for PCP, heroin, and use of needles with heroin. Finally, some frequency of use data is presented for marijuana, cocaine, and alcohol.

Table 2.5:
Number of People Using Illicit Drugs[45]
(in thousands)

	Ever Used	Used in Past Year	Used in Past Month
Total	72,496	27,971	14,479
Age 12-17	5,005	3,404	1,866
Age 18-25	17,491	9,485	5,290
Age 26-34	24,768	8,730	5,008
Age 35+	25,232	6,351	2,316
White	58,041	21,783	10,936
Black	7,999	2,966	1,734
Hispanic	4,823	2,189	1,218

[44]See Hosek and Overberg (1989), Ryan and Miller (1989), and *International Drug Report* (1989b).

[45]U.S. Department of Health and Human Services, 1989c, Tables 2-A, 2-B, 2-C, 2-D, pp.17-19.

The most striking thing about the National Household Survey Data is that the numbers are large. Many people use or have used drugs. (See Table 2.5.)

Besides showing that many people use illicit drugs, Table 2.5 shows that, contrary to the image projected by the media, the vast majority of users are white. Of course one reason for this is that most Americans are white, but as Table 2.6 shows, a larger fraction of white Americans use drugs than do either blacks or Hispanics.

Table 2.6:
Prevalence of Drug Use by Different Race/Ethnic Backgrounds[46]

	White	Black	Hispanic
Total			
Ever Used	37.0%	35.9%	32.3%
Past Year	13.9	13.3	14.7
Past Month	7.0	7.8	8.2
	White	Black	Hispanic
Age 12-17			
Ever Used	26.0	18.7	24.3
Past Year	17.8	12.1	16.3
Past Month	10.0	6.2	7.3
Age 18-25			
Ever Used	62.5	47.0	47.6
Past Year	33.1	25.9	28.7
Past Month	18.0	16.9	16.8
Age 26-34			
Ever Used	67.4	58.0	50.9
Past Year	22.8	21.8	19.8
Past Month	13.3	11.2	11.8
Age 35+			
Ever Used	22.8	27.0	17.1
Past Year	5.7	5.2	4.4
Past Month	1.8	3.3	2.2

Why, one might ask, do the media so frequently portray drug use as a predominantly inner-city, minority problem? There are several possible answers short of racist conspiracies. One is that blacks are more likely to use relatively more harmful drugs.

[46]U.S. Department of Health and Human Services, 1989c, Tables 2-A, 2-B, 2-C, 2-D, pp.17-19.

Prevalence is higher for whites for marijuana, inhalants, hallucinogens, PCP, and most psychotherapeutics. Prevalence is higher for blacks for cocaine (except the ever used category), crack, heroin, and intravenous drug use. Also, the media tends to focus on cocaine, especially crack.

The frequency of use data from the National Household Survey are scanty, but they suggest that this may also be a factor. Specifically, it might be that many whites experiment with drugs, but a smaller fraction go on to regular use.

It is also possible that whites generally have access to better medical services and support groups. More white users may be treated or counseled early in their using career, before they begin to do things that attract public attention. Poor blacks, in contrast, may, out of poverty, be more likely to turn to street crime to finance their habit.

Finally, much of the violence associated with the drug trade occurs in the inner city, and that is where blacks in general and black drug users in particular, are concentrated.

2.5.2 The High School Senior Survey

Since 1975 NIDA has also sponsored a survey of young people. The survey is commonly referred to as the "High School Senior Survey"[47] although it now reports on use by college students and young adults as well. Each year about 17,000 high school seniors are questioned. Of that group, about 2,400 are selected for follow up. These 2,400 are divided into two groups, each of which is surveyed by mail in alternating years.

The most significant bias in the survey is that it omits high school drop outs. About 15 - 20% of each class drop out before the survey is administered.

Unlike the National Household Survey, the High School Senior Survey only covers a small fraction of the population. It has the advantages, however, of being conducted every year and including more detailed information about frequency of use, degree and duration of highs, perceived harmfulness, social disapproval, proportion of friends using, and perceived availability of drugs.

It is interesting to look at how various quantities estimated by the High School Senior Survey changed over time. The simple answer is, they have been falling. For most subgroups, most drugs,

---

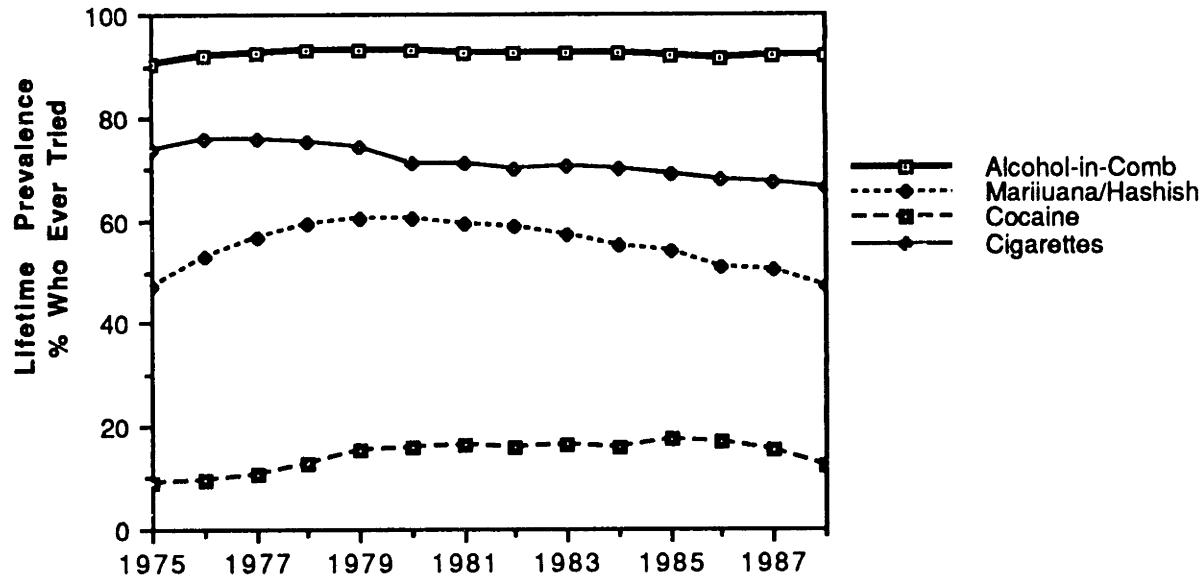
[47]Information in this subsection is taken from U.S. Department of Health and Human Services, *Illicit Drug Use, Smoking, and Drinking by America's High School Students, College Students, and Young Adults: 1975-1987*, 1988c.

and most measures, drug use has been falling steadily and substantially among those surveyed for the last five to ten years.

Figures 2.8a[48] and 2.8b[49] show the trend in lifetime prevalence for high school seniors between 1975 and 1988.[50] The figures for alcohol are high and stable. Those for cigarettes increased slowly until 1986 and have fallen since. In contrast, cocaine use peaked around 1980 and has been falling since. Marijuana use, although still very common, has been falling quite steadily throughout the study period. Declines in the use of hallucinogens, barbituates, methaqualone, and tranquilizers were even more dramatic. Inhalants are the only class of drug for which prevalence rates have risen.

Figure 2.9[51] shows parallel data for college students. The trends are generally similar, although levels sometimes differ between the two populations.

Figure 2.8a
High School Senior Survey Trends in
Lifetime Prevalence Percent Who Ever Tried



[48]Data from U.S. Department of Health and Human Services, 1989a.

[49]Data from U.S. Department of Health and Human Services, 1989a.

[50]Isikoff (1990) reports that the general downward trend continued in the 1989 data.

[51]Data taken from U.S. Department of Health and Human Services, 1989b, "NIDA Capsules: College Students Survey on Drug Abuse: 1980-1988."

Figure 2.8b
High School Senior Survey Trends in Lifetime Prevalence, #2

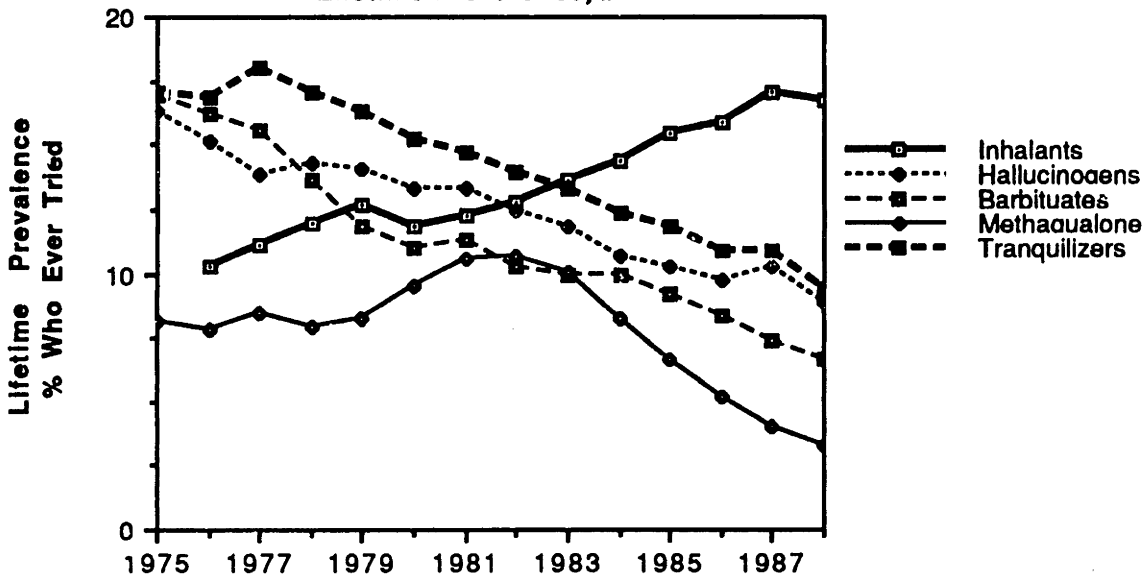
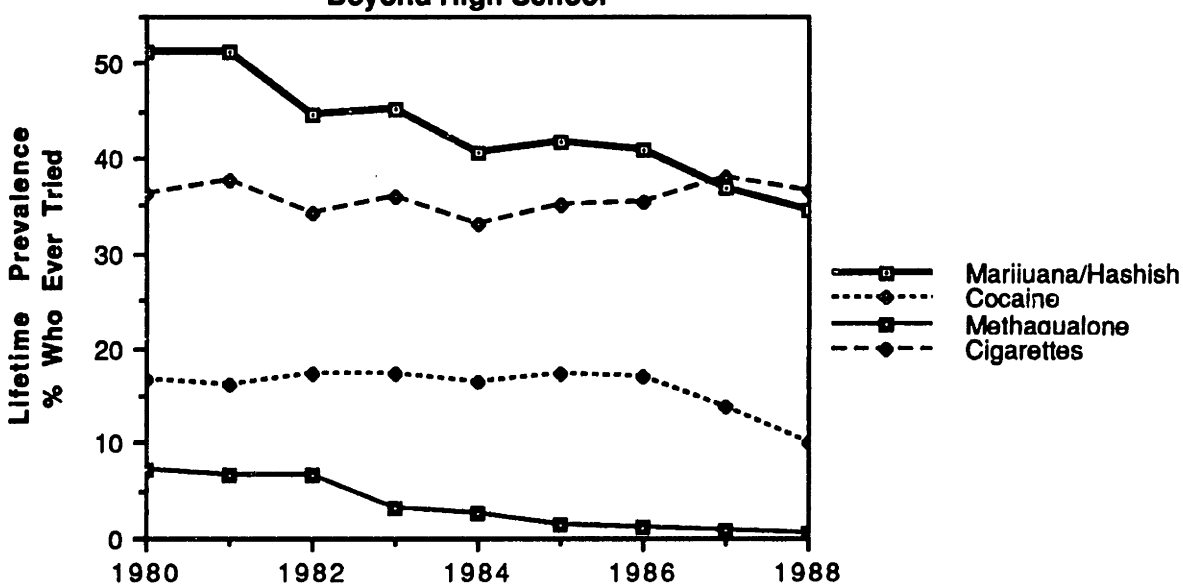


Figure 2.9
Trends in Annual Prevalence of Drug Use Among College Students 1-4 Years Beyond High School



The overall trend is clear and surprising. Prevalence of drug use appears to be declining among young people. This does not necessarily mean that drug use on the whole is declining for at least

two reasons. First, trends in the study group may not be replicated in the larger population. Second, even if fewer people are trying drugs, those who do so might be using larger quantities.

2.5.3 The Drug Abuse Warning Network (DAWN)

NIDA has also been collecting data about drug-related emergency room and medical examiner episodes since 1972. Every year it publishes books of tables breaking down the results for various demographic factors and characteristics of the episode.[52]

The data do not cover the entire population, nor do they constitute a random sample. In 1987, "a total of 756 hospital emergency rooms reporting 146,778 episodes and 75 medical examiner facilities reporting 4,678 deaths were affiliated with the DAWN system."[53] These facilities are a nonrandom sample of those in 27 metropolitan areas which include roughly one-third of the U.S. population.

Extrapolating to national levels is discouraged, but since there are few if any alternatives, such extrapolations are made. For example, Kleiman roughly approximates national numbers of emergency room mentions for marijuana by multiplying the DAWN number by 3.5. This factor is based on the facts that the cities surveyed in 1986, the year for which he made his estimate, included 29% of the U.S. population, 76% of the hospitals in those areas participated in the survey, and there are about 30% fewer marijuana users in nonmetropolitan areas.[54]

Such rough calculations are probably reasonable. Getting more precise is difficult for a variety of reasons. For example, in 1987 medical examiner data was not available for New York City, there were problems with the reporting by an emergency room in Miami, the medical examiners in Boston had difficulty reporting their total case load, and the original 1972 metropolitan area boundaries are still used to facilitate trend analysis.

DAWN collects information on "patient (or decedent) sex, race, and age; drug concomitance; drug use motive (or manner of death); reason for ER contact (or cause of death); disposition of ER patient; source of substance (ER reports only); form in which drug was

---

[52]One such book is *Annual Data 1987: Data from the Drug Abuse Network (DAWN)*, U.S. Department of Health and Human Services, 1988d. Unless otherwise noted, it is the source of all information about DAWN presented in this subsection.

[53]U.S. Department of Health and Human Services, 1988d, p.1.

[54]Kleiman, 1989, p.12.

acquired or found; route of drug administration; and location of facility within the metropolitan area (ER reports only)."[55]

Information about the source of drugs would seem to be very valuable, but the categories are broad (Legal prescription, street buy, other unauthorized procurement, other, and unknown), and for most drugs about half of the responses are "unknown".

Despite its weaknesses the DAWN data are useful. For instance, Tables 2.7 and 2.8 give a sense of the magnitude of the health effects of the drug problem. Note, more than one drug may be "mentioned" in a single "episode". For example, someone who smoked a combination of heroin and cocaine after drinking would produce three mentions, one each for heroin, cocaine, and alcohol-in-combination.

It probably comes as no surprise to the reader that cocaine and heroin are among the three most frequently listed drugs for both emergency room and medical examiner mentions. What might be more surprising is that alcohol-in-combination rounds out the top three in both cases.

The other drugs individually account for far smaller shares of all mentions. Marijuana is rarely mentioned by medical examiners (and then only in combination with other drugs), and considering how many people use it, the number of emergency room mentions is not that high.[56]

Table 2.7:
Ten Drugs Mentioned Most Frequently In Emergency Room Episodes
DAWN Data for 1987[57]

Drug Name	Number of Mentions	Percent of Total Episodes
Cocaine	46,331	31.57
Alcohol-in-combination	40,644	27.69
Heroin/Morphine	18,566	12.65
Marijuana/Hashish	10,083	6.87
PCP/PCP Combination	9,545	6.50
Diazepam	7,277	4.96
Acetaminophene	6,469	4.41
Aspirin	5,688	3.88
Alprazolam	3,835	2.61
Ibuprofen	3,612	2.46

[55]U.S. Department of Health and Human Services, 1988d, p.7.
[56]Kleiman (1989, pp.11-13) describes this in detail.

Table 2.8:
Drugs Mentioned by Medical Examiners in More Than 5% of Episodes
DAWN Data for 1987[58]

Drug Name	Number of Mentions	Percent of Total Episodes
Alcohol-in-combination	1,730	36.98
Cocaine	1,696	36.25
Heroin/Morphine	1,572	33.60
Codeine	590	12.61
Quinine	403	8.61
Diazepam	312	6.67
Amitriptyline	299	6.39
PCP/PCP Combinations	249	5.32
D-Propoxyphene	241	5.15

The DAWN system yields other insights. For example, of the 4,678 deaths reported, 1,283 (27.4%) were suicides. For another 586 (12.5%) the manner of death was unknown. Only 2,809 (60.0%) were listed as accidental/unexpected.[59] Of course this is still a large number, but one must remember that a significant fraction of the deaths and overdoses not leading to deaths are the results of suicides and attempted suicides. If illicit drugs were to suddenly cease to be available, it might well be that fewer lives would be saved than one might initially expect because those who were trying to end their own lives might find other means of doing so.

A second point is simply that the death toll directly attributable to drug use is not as great as one might think. Even if the rest of the country had as many deaths per capita as these metropolitan areas, there would still have been "only" 15,000 or so deaths directly attributable to illicit drug use. The figures cited for alcohol (100,000) and cigarettes (500,000) are much higher,[60] although they probably include many less direct causes of death and so may not be comparable.

The DAWN data excludes deaths from AIDS spread by IV drug use, motor vehicle and other transportation accidents caused by drug

---

[57]Extracted from U.S. Department of Health and Human Services, 1988d, Table 2.06a, p.26.

[58]Extracted from U.S. Department of Health and Human Services, 1988d, Table 3.06a, p.53.

[59]U.S. Department of Health and Human Services, 1988d, p.51.

[60]These figures are from *Drug Abuse Update*, No. 26, September, 1988, published by The National Drug Information Center of Families in Action and The Scott Newman Center.

use, premature deaths caused by side effects of drug use such as malnutrition and poverty, and drug related homicides. Furthermore, many victims of drug abuse suffer even if they do not die. Many people, children as well as adults, suffer from addiction. Nationally perhaps 30,000-50,000 crack-babies are born each year,[61] and many of these children may never fully recover emotionally, intellectually, or physically.[62]

The DAWN data also clearly show there are substantial regional variations in drug use. For example, the number of mentions per 1,000 deaths reported by medical examiners varies widely from city to city. (See Table 2.9.)

Table 2.9:
Drug Mentions per 1,000 Deaths from all Causes
DAWN Medical Examiner Data for 1987[63]

Drug Name	San Diego	Washington D.C.	All DAWN Metropolitan Areas
Heroin/Morphine	24.4	60.0	13.5
Cocaine	14.6	48.7	14.4
Codeine	10.9	1.1	4.9
PCP/PCP Combinations	1.9	29.3	2.2

Figure 2.10a, 2.10b, and 2.11 show trends in DAWN data. Trend data are based on cases reported by 564 hospital emergency rooms and 62 medical examiners facilities that reported consistently between 1976 and 1985.[64]

Trends in emergency room admissions (Figures 2.10a[65] and 2.10b[66]) vary between drugs. Episodes involving diazepam (Valium) have declined markedly. Heroin episodes declined until 1979, but have been rising since then. The number of episodes involving most other drugs have been rising, with cocaine related episodes growing substantially from 1,015 in 1976 to 9,403 in 1985.

---

[61]Besharov, 1989.

[62]Kantrowitz et al., 1990.

[63]Extracted from U.S. Department of Health and Human Services, 1988d, Table I-6, p.236.

[64]U.S. Department of Health and Human Services, *Trends in Drug Abuse Related Hospital Emergency Room Episodes and Medical Examiner Cases for Selected Drugs: DAWN 1976-1985*, 1987b.

[65]U.S. Department of Health and Human Services, 1987b, Table B.1, p.283.

[66]U.S. Department of Health and Human Services, 1987b, Table B.1, p.283.

Figure 2.10a
Number of Emergency Room Episodes Involving Alcohol-in-Combination, Diazepam, Heroin/Morphine, and Aspirin

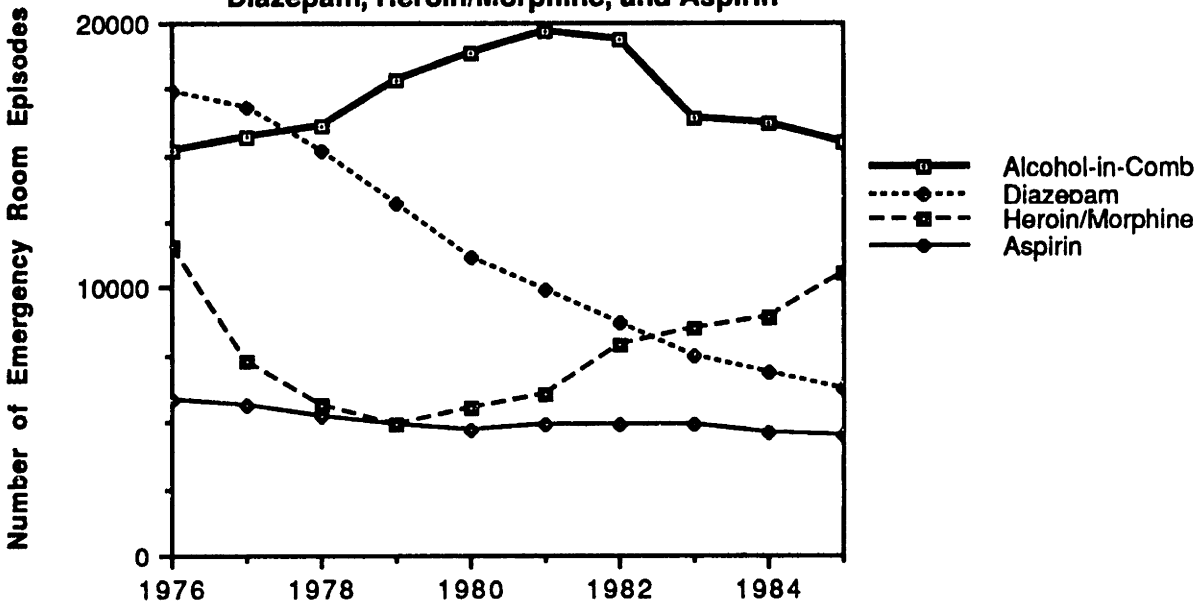
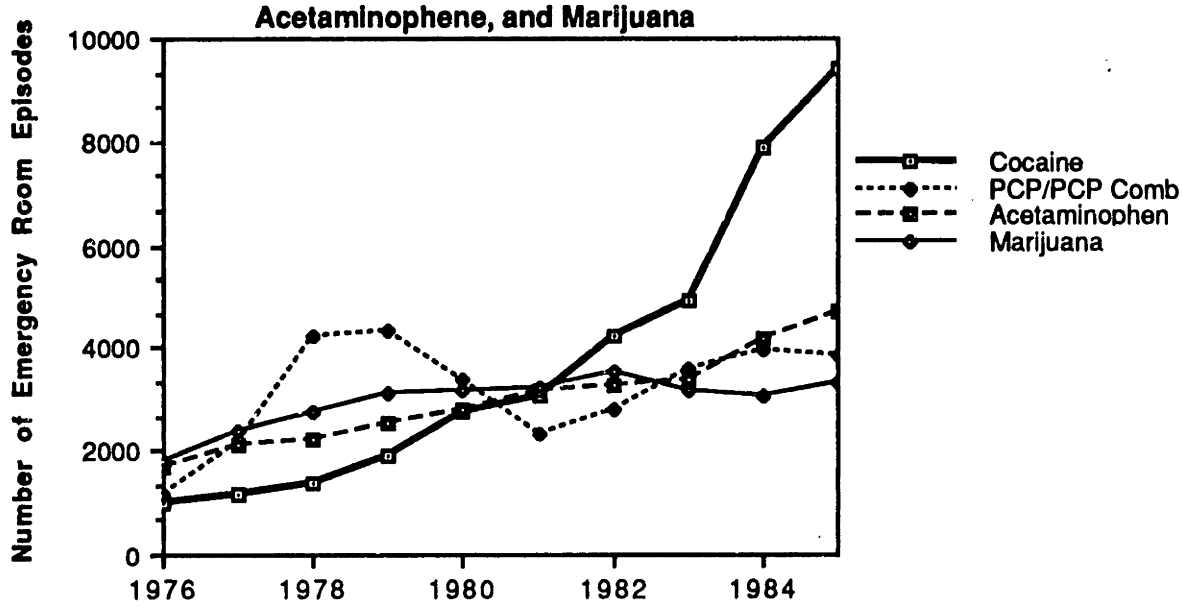
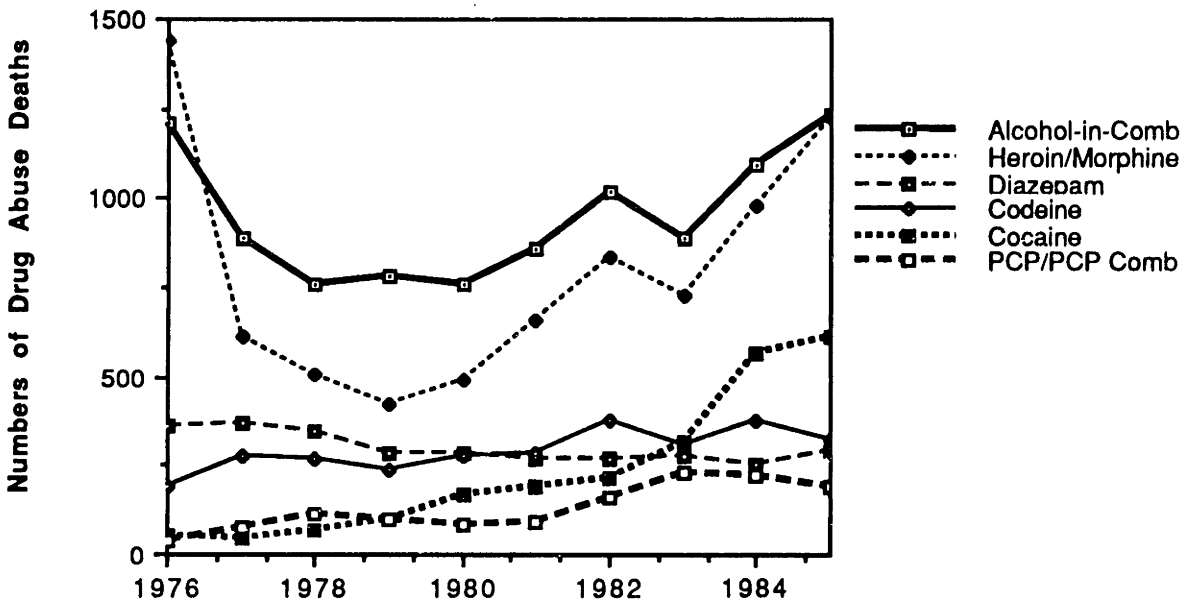


Figure 2.10b
Number of Emergency Room Episodes Involving Cocaine, PCP/PCP Cominations, Acetaminophene, and Marijuana



Drug abuse deaths reported have followed similar patterns. (See Figure 2.11.[67]) The number of heroin related deaths fell until 1979, but has rebounded. Cocaine related deaths rose from 53 in 1976 to 615 in 1985. Deaths from most other drugs except diazepam are up, but not nearly so dramatically.

Figure 2.11
Number of Drug Abuse Deaths Involving the Six Most Commonly Cited Drugs



Hence it appears that although the number of people using drugs is declining, the number who are using to excess or using in dangerous ways is increasing. This is not implausible; only a small fraction of people who have ever used drugs end up in emergency rooms or morgues because of it. So one should not be surprised if the total number of users and the DAWN data are not highly correlated.

2.5.4 Data on Clients Admitted to Drug Treatment Programs

NIDA also reports trends in characteristics of clients admitted to State-monitored drug abuse treatment programs. Its 1988 report, *Demographic Characteristics and Patterns of Drug Use of Clients Admitted to Drug Abuse Treatment Programs in Selected States:*

---

[67]U.S. Department of Health and Human Services, 1987b, Table B.4, p.287.

Trend Data 1979-1984, includes data from a panel of 596 programs in 15 states which consistently reported to NIDA.

The body of the report devotes a chapter to each of the following drugs: alcohol, amphetamines, barbituates, cocaine, heroin, marijuana, opiates other than heroin, PCP, and tranquilizers. Each chapter gives annual data for information such as the number of client admissions, sex, race/ethnic background, age at admission, employment status, last formal school year completed, frequency of use, route of administration, prior drug treatment experience, years between first use and first treatment admission, age at first use, source of referral, and secondary drug use.

An additional chapter gives some information about clients admitted for problems with inhalants, hallucinogens other than PCP, and sedatives other than barbituates.

The report gives some indication about trends in the demographic composition of the user population. Neither the statistics nor the trends reported can be directly extrapolated to the entire population of users, however, because those seeking treatment may be a biased subsample of the larger population. For example, they might tend to be older, have used longer, and to be using more dangerous drugs. Also, the programs reporting come from only 15 states, so regional variations may bias the sample. Finally, trends in client admissions may lag trends in the general population; if PCP were to suddenly become more popular in 1991, that probably would not affect the composition of those seeking treatment until 1992 or later.

Nevertheless some trends one might expect to see are apparent. For example, people seeking treatment for problems with cocaine in recent years are more likely than before to be black, female, and to have used cocaine by smoking. This probably reflects the spread of crack. Likewise, the average age at admission for those seeking treatment for heroin has increased steadily, probably reflecting the aging of the heroin using population as a whole. So the data might at least be useful for suggesting hypotheses about changes in the user population that one could seek to confirm independently.

2.5.5 The Drug Use Forecasting System (DUF)

Since 1986 data on drug use by arrestees have been collected by the Drug Use Forecasting (DUF) System. The program was initiated (with NIJ funding) by Narcotic and Drug Research, Inc. when it obtained voluntary interviews and urine samples from male arrestees in Manhattan's Central Booking Facility. A follow-up study

was conducted in 1986.[68] It was expanded to about 10 cities in 1987 and to 22 in 1989.[69] It has since begun producing information about subjects such as crack distribution and consumption.[70]

The program obtains voluntary interviews and urine samples from about 250 arrestees in each city every three months. Prevalence data are based on the urine samples. It is intended to track trends in drug use by urban criminals; it has little to say about trends in the larger population.

The most striking result of these studies is that a very large fraction of urban arrestees use illicit drugs. In the January-March 1989 sample taken from 13 cities, the percentage male arrestees testing positive for some drug (including marijuana) ranged from a low of 50% in Indianapolis to a high of 85% in San Diego. Most were between 60% and 80%. Among the same sample the percentage of arrestees testing positive for cocaine ranged from a low of 24% in San Antonio to a high of 76% in New York City. Most of those values were between 40% and 60%.[71]

2.6 Data On Prices

2.6.1 Introduction to the Price Data

Prices generally reveal a great deal about markets, and the data on drug markets are no exception. Perhaps the most striking thing the price data show is that at the retail level, illicit drugs are fantastically expensive. (See Table 2.10.)

---

[68]Wish (1987) describes the initial studies.

[69]The expanded program is described by Wish and O'Neill (1989) and Stewart (1988).

[70]See Mieczkowski, 1989.

[71]Wish and O'Neill, 1989.

Table 2.10:
Current Retail Prices of Various Commodities
(Dollars per kilogram)

Commodity	Price
Pure Heroin (retail)[72]	$1,000,000
Pure Cocaine (retail)[73]	$100,000
THC (retail)[74]	$100,000
Gold[75]	$13,200
Commercial Grade Marijuana (retail)[76]	$4,300
Silver[77]	$135
Mercedes 560SL Convertible[78]	$36
Steak[79]	$15
Sugar[80]	$0.88
Flour[81]	$0.44

One might well wonder how slightly processed agricultural products (such as heroin, cocaine, and marijuana) can be so expensive. A simple answer is that although enforcement does not seem to be able to make drugs unavailable, it can make them expensive. Table 2.11 shows how drugs' value increases as they move down the distribution chain. The data are not recent, but the basic price structure has not changed in the last ten years.

---

[72]Based on a point representation of the range of retail prices given by the DEA's Office of Intelligence April-June 1989 report and an assumed retail purity of 20%.

[73]See Table 3.6

[74]See Table 3.8.

[75]Gold (Comex) spot of $374.30/oz., *The New York Times*, April 12, 1990.

[76]Table 2.13, midpoint of range.

[77]Silver (Comex) spot of $5.129/Troy oz., *The New York Times*, April 12, 1990.

[78]From the Consumer Guide Automobile Book, 1990 Edition. Based on a retail price of $64,230 and a curb weight of 3,970 pounds.

[79]$6.79/lb. at Stop and Shop super market in Cambridge, MA on April 13, 1990.

[80]$0.398/lb. at Stop and Shop super market in Cambridge, MA on April 13, 1990.

[81]$0.198/lb. at Stop and Shop super market in Cambridge, MA on April 13, 1990.

Table 2.11:[82]
Structure of Drug Prices, 1980
(per Pure Kilogram)

	Heroin	Cocaine	Marijuana[1]
Farmgate	$350-1000[2]	$1,300-$10,000	$7-$18
Processed	$6,000-$10,000	$3,000-$10,000	$55
Export	$95,000	$7,000-$20,000	$90-$180
Import[3]	$220,000-$240,000	$50,000	$365-$720
Retail	$1.6-$2.2 million	$650,0000[4]	$1,250-$2,090

1 Prices are for Colombian-origin marijuana, estimated to account for 75 percent of total U.S. consumption in 1980.
2 The price of the 10 kg of opium required to manufacture 1 kg of heroin.
3 The import price refers to the price at first transaction within the United States. Marijuana is purchased in roughly ton lots, cocaine in multikilo lots, and heroin in kilo lots.
4 The original data source reported a retail price of $800,000. Other DEA data, such as those reported in U.S. General Accounting Office (1983), consistently indicate prices in the range of $600,000-$650,000 in 1980.

These data show that there is not just one market for each drug. There are separate markets depending on the location and the quantity traded. For example, a kilogram of cocaine or heroin ready to be shipped to the U.S. is worth less than half as much as the same kilogram on the beach in the U.S.. Clearly cocaine and heroin traded in source countries is not in the same market as cocaine and heroin traded here.

Even within the U.S. there are several markets. The kilogram on the beach is worth a fraction of its value broken down into retail quantities and sold on the street. Cocaine in a kilogram brick is simply not the same commodity as cocaine ready to be sold at retail.[83] The domestic distribution network may not do much to change the drugs physically, but considerable value is added simply by moving them off the beach and into retailers' hands.

Actually drugs do not move directly from the importer to the retailers. They pass through a chain of intermediaries (wholesalers), which are referred to collectively in this thesis as the domestic drug distribution network. Layered distribution chains exist for licit

---

[82]Table (including notes) copied from Reuter and Kleiman (1986, p.293). They quote as a source "Adapted from National Narcotics Intelligence Consumers Committee (1982)." Wiant (1985) gives a slightly more detailed description of the heroin price chain.
[83]Reuter, 1988.

goods as well. Generally the factory is at the highest level; the regional warehouses at the second level; local warehouses at the third; and finally, the retail stores comprise the lowest level of the distribution network.

As with licit goods, prices rise and transaction quantities fall as one moves down the network. The principal differences are the magnitude of the price increases and the lack of central control. With licit goods one company frequently owns the factory, the warehouses, and sometimes even the retail stores. With illicit drugs, although some organization have maintained a degree of vertical integration, participants at different levels are usually in business for themselves or at least are not working for the same organization.

Hence a brief look at the price data support two broad conclusions. The first is simply that illicit drugs are extremely expensive. The second is that there are distinct markets for the drugs depending on their location and the size of the transaction. The individuals and organizations participating in the markets form a distribution chain, not entirely unlike those that exist for licit goods. The part of the chain between the importers and retailers will be referred to here as the domestic distribution network.

2.6.2 Wholesale and Retail Price Data

The next chapter examines how changes in prices at one level of the domestic distribution network affect prices at subsequent levels. It discusses two theories that make two distinct predictions and compares these theories' predictions with historical data. Ideally one would observe changes in import and retail prices because the price differences are greatest between those two levels. However, few import data are available, so wholesale and retail data will be used instead.[84] This section presents the raw data that will be used.

The Drug Enforcement Administration's Office of Domestic Intelligence regularly collects data on wholesale and retail prices for a variety of drugs. As will be discussed in Section 3.9.3, at present only the cocaine and marijuana data can be used to test the theories in Chapter 3. Tables 2.12 and 2.13 compile cocaine and marijuana data, respectively, from a variety of DEA reports.[85]

---

[84]Reuter, Crawford, and Cave (1988) used wholesale data in their study for the same reason.

[85]The author is indebted to Steve Morreale and Maurice Rinfret of the DEA for their assistance in obtaining these data.

Table 2.12:[86]
Cocaine Prices, 1982 - 1989

	1982	1983	1984	1985
Wholesale (1 kg)				
National	$55,000-65,000	$45,000-55,000	$40,000-50,000	$30,000-50,000
Miami	47,000-60,000	25,000-30,000	24,000-28,000	28,000-37,000
New York	50,000-65,000	35,000-45,000	30,000-45,000	34,000-40,000
Chicago	55,000-70,000	45,000-55,000	35,000-55,000	40,000-45,000
Los Angeles	55,000-70,000	45,000-55,000	30,000-45,000	35,000-40,000
Wholesale/Retail (1 oz)				
National	$2,000-2,800	$1,800-2,500	$1,800-2,400	$1,600-2,300
Miami	2,000-2,400	1,600-2,200	1,600-2,000	1,200-1,600
New York	2,000-2,600	1,600-2,400	1,600-2,200	1,400-2,000
Chicago	2,000-2,800	1,700-2,400	1,700-2,400	1,600-2,300
Los Angeles	2,000-3,000	1,600-2,500	1,600-2,400	1,500-2,000
Retail (1 gm)				
National	$100-125	$100-125	$100-120	$100
Miami	100	50-90	50-90	50-70
New York	100	75-100	75-100	75-100
Chicago	100-125	100	100	100
Los Angeles	100-150	100	100	100

	1986	1987	1988	1989(Apr-Jun)
Wholesale (1 kg)				
National	$22,000-45,000	$12,000-40,000	$11,000-34,000	$9,000-32,000
Miami	15,000-25,000	12,000-15,000	13,000-20,000	12,500-18,000
New York	18,000-28,000	15,000-30,000	16,000-23,000	15,000-23,000
Chicago	30,000-45,000	20,000-40,000	17,000-24,000	14,000-21,000
Los Angeles	25,000-35,000	10,000-18,000	11,000-16,000	12,000-15,000
Wholesale/Retail (1 oz)				
National	$1,300-2,300	$800-2,100	$500-2,000	$500-2,000
Miami	800-1,200	800-1,200	800-1,200	800-1,200
New York	1,200-1,800	800-1,600	600-1,000	600-1,000
Chicago	1,500-2,000	1,100-1,800	750-1,400	700-1,100
Los Angeles	1,5000-2,200	600-1,000	500-800	500-800
Retail (1 gm)				
National	$80-120	$80-120	$50-120	$50-125
Miami	50-60	50-60	55-85	50-80
New York	70-100	80-100	50-90	50-80
Chicago	100	100	75-100	70-100
Los Angeles	100	100	50-100	70-110

[86]DEA Office of Intelligence Illicit Drug Wholesale and Retail Price Reports September 1984, October - December 1988, and April - June 1989.

Table 2.13:
Marijuana Prices, 1982-1989[87]

	Wholesale ($/lb.)	Retail ($/oz.)	Purity
Sinsimella			
1982	$1,000-$2,000	$100-$125	
1983	$1,000-$3,000	$100-$150	
1984	$1,200-$2,500	$120-$180	6.73%
1985	$1,200-$2,000	$120-$200	7.28%
1986	$800-$2000	$100-$200	8.44%
1987	$1,400-$2,100	$160-$210	7.97%
1988	$800-$3000	$120-$300	8.43%
1989(Apr-Jun)	$800-$4000	$120-$300	7.91%
Commercial Grade			
1982 Domestic	$350-$600	$40-$50	
Mexican	$350-$550	$40-$50	
Jamaican	$400-$600	$45-$65	
Colombian	$350-$500	$30-$40	
1983 Domestic	$350-$650	$40-$65	
Mexican	$350-$600	$40-$60	
Jamaican	$450-$600	$45-$65	
Colombian	$350-$1,200	$60-$175	
1984	$400-$600	$45-$75	3.78%
1985	$300-$600	$50-$100	3.70%
1986	$350-$700	$45-$120	3.34%
1987	$350-$1450	$30-$250	3.46%
1988	$350-$1800	$30-$250	3.63%
1989(Apr-Jun)	$350-$2000	$20-$225	3.79%

One can readily see from the tables that one of the difficulties with working with price data is that there is no one number for the retail or the wholesale price. The data are divided by region and more importantly are almost always given as a range, and the range can be quite broad. Furthermore, it is not clear whether the range is a confidence interval, the actual observed range, or something else (such as the observed range trimmed of outliers).

2.7 "Case Studies" of Drug Markets and Participants

The previous section described sources of data and information about the aggregated national market, studies of the "macro" drug

---

[87]DEA Office of Intelligence Illicit Drug Wholesale and Retail Price Reports September 1984, October - December 1988, and April - June 1989.

economy. The "case study approach" to drug markets has also been used. These in depth studies and interviews of a relatively few people can give a better understanding of how the market works than does just describing its aggregate characteristics. Their chief weakness is that there is tremendous heterogeneity among both consumers (users) and produces (dealers).

No attempt is made here to summarize these studies, and unlike the list of data sources above, the set of studies mentioned here attempts only to illustrate the genre, not to comprise an exhaustive list. The studies are divided into two groups: formal studies conducted by trained researchers and journalistic accounts.

2.7.1 Formal Ethnographic Studies

Some of the leading ethnographic research on drug markets comes from a school of research founded by Edward Preble and John J. Casey.[88] This group of researchers studied the lives of inner-city people involved with drugs, notably heroin, paying particular attention to the economic environment in which they live. Johnson et al.[89] is a relatively recent and very significant product of this work. It contains detailed information about the daily incomes, expenditures, and drug consumption of a sample of 201 street addicts.

Studying retail markets through interviews and direct observation is challenging, but studying higher-level markets may be even more difficult, if for no other reason than it is hard to meet high-level dealers. Reuter and Haaga[90] circumvented this difficulty by interviewing 40 drug offenders in federal prisons.

Their study was a pilot-project to assess the feasibility of gathering information by talking to people in prison. Low and selective participation, however, made them question whether a larger study would be desirable. Nevertheless, the study did yield some interesting results.

Adler[91] is the only researcher of whom the author is aware who was able to study high-level dealers in operation. Her book is a valuable source of information about the lifestyle, customs, and attitudes of one group of white, wholesale dealers in southern California. The extent to which the group she studied is representative of all dealers is hard to judge.

---

[88]Preble and Casey, 1969.

[89]Johnson, et al., 1985.

[90]Reuter and Haaga, 1989.

[91]Adler, 1985.

Finally, Reuter and Haaga mention three other sociological studies in their review of prior research on markets and dealers' careers. They are John Langer's "Drug Entrepreneurs and Dealing Culture" (*Social Problems*, Vol. 24, 1977, pp.377-386), John Leib and Sheldon Olson's "Prestige, Paranoia, and Profit: On Becoming a Dealer of Illicit Drugs in a University Community" (*Journal of Drug Issues*, Vol. 6, 1975), and Lawrence Redlinger's "Marketing and Distributing Heroin" (*Journal of Psychedelic Drugs*, Vol. 7, 1975, pp.331-353).

2.7.2 Journalistic Accounts

Journalists have also written about drug dealers. On the one hand most of these authors are not trained social scientists, and frequently their accounts are based on one or at most a few people. On the other hand, if they are honest, accurate, and can convey what they observe in writing, these authors can be a valuable source of information.

One of the best known books in this genre is Richard Woodley's *Dealer: Portrait of a Cocaine Merchant* which is about a low-level cocaine dealer in Haarlem in the early 1970's. James Mills' *The Underground Empire: Where Crime and Governments Embrace* provides a glimpse of three targets of DEA CENTAC (Central Tactical Unit) operations in the late 1970's and early 1980's. Reuter and Haaga's literature review also mentions a book by Roger Warner (*Invisible Hand: The Marijuana Business*, Beech Tree Books: New York, 1986) and a "semi-fictional" account of high-level dealers by Robert Sabbag (*Snowblind*, Avon: New York, 1978).

2.8 Unconventional and Indirect Data Sources

Since the standard data sources leave something to be desired, drug market researchers sometimes rely on nonstandard sources for clues. A fairly well-known example of this is that one piece of evidence supporting the belief that Florida was a major import site for illicit drugs was the large cash surplus at the Miami Federal Reserve Bank. Most Federal Reserve banks received a little less cash than they paid out. In Miami there was a substantial cash surplus. More recently the Los Angeles Federal Reserve Bank has begun running large cash surpluses, and it is generally believed that some smuggling has shifted to the west coast because interdiction rates on the east coast increased.

This section briefly discusses two indirect sources of information about drug markets. The second, sales of rolling paper, is a relatively unknown source.

2.8.1 The HIV/AIDS Surveillance Report

Every month the Center of Disease Control issues a bulletin giving information about the number of AIDS cases reported in the most recent year, the preceding year, and the cumulative total.[92] The report includes tables identifying the exposure category and demographic characteristics.

The cumulative number of AIDS cases reported through December 1989 for drug use related exposure categories were: female and heterosexual male intravenous (IV) drug user (24,212), male homosexual/bisexual intravenous drug user (8,117), heterosexual sex with an IV drug user (2,871), pediatric case for which mother was an IV drug user (826), and mother had sex with an IV drug user (330).

AIDS is an important part of "the drug problem," and these reports give information directly about that part of the problem. The reports give relatively little useful information about drug use in general, however, because not enough is known about the fraction of IV drug users have AIDS. Also, in as much as AIDS is a contagious disease and there is not random mixing within at-risk populations, there is every reason to believe that IV drug users who have AIDS are not a random sample of IV drug users in general.

2.8.2 Rolling Paper Consumption

The Bureau of Alcohol, Tobacco, and Firearms (BATF) collects taxes on the paper produced to hand-roll cigarettes. They have data on how much tax was collected each year since 1975 and each month since October, 1982. This subsection describes how this information might be used to estimate marijuana consumption.

If one believes that:

Assumption 1: The vast majority of rolling papers are used to roll marijuana joints, and

Assumption 2: Most marijuana joints are made with rolling paper

then if ones knew how many rolling papers were used each year, one would also have a good estimate of the number of marijuana joints produced (and hence presumably consumed) each year.

---

[92]For example, Centers for Disease Control, January, 1990 and April, 1990.

Assumption 2 is certainly plausible. Rolling paper is legal and inexpensive, and regular paper cannot be used, so marijuana users have little incentive not to use rolling paper.

Furthermore, if one believes that

Assumption 3: Most marijuana consumed is smoked in joints,

then this also allows one to estimate total marijuana consumption. It seems likely that Assumption 3 is true.

Hence estimating marijuana consumption from rolling paper data hinges on Assumption 1. It seems reasonable that most rolling papers are used for either tobacco or marijuana. Most rolling paper is probably purchased with the intention of making a cigarette of some kind, and Americans smoke much more marijuana and tobacco than any other substance. It is possible that a significant fraction of rolling papers purchased are never used. They are so inexpensive one would not expect users to steward them they way they would their drugs. On the other hand, the papers are not perishable, so it is at least plausible that most are eventually used.

So the real question about Assumption 1 is whether or not tobacco cigarettes account for a significant fraction of all hand-rolled cigarettes. Clearly the vast majority of tobacco cigarettes smoked in the U.S. today are not rolled by hand. However, there are many more tobacco cigarettes smoked in the U.S. than marijuana joints (roughly 600 billion cigarettes versus 12 billion marijuana joints[93]) so even if only 1% of tobacco cigarettes were hand-rolled, they would account for 33% of all rolling papers used.

The assumption here is that the fraction of all tobacco cigarettes that are hand rolled is negligible (say less than 0.1%) so the vast majority of hand-rolled cigarettes are marijuana joints.

The next question is whether one can determine from BATF data how many rolling papers are consumed. For a variety of reasons this is not as easy as it might seem. For one, the BATF only taxes rolling papers sold in books of 25 or more papers.[94] It is not known what fraction of rolling papers are sold in smaller sets. Also, the annual data reports (years 1976-1982) report the number of books sold containing 30, 50, 75, 100, 150, 300, and in one case 200 papers each. The more recent monthly data also include categories for books with 25, 32, 36, 45, 48, 64, 66, 72, and 78 papers per set.

---

[93]Kleiman, 1989, p.8.

[94]Bureau of Tobacco, Alcohol, and Firearms, 27 CFR Ch. 1, Section 285, April 1, 1988.

It seems possible that the earlier data were "rounded" in the sense that sales of 72 paper sets were reported as sales of 75 paper sets. Finally, before 1982 the BATF reports data on "cigarette paper cut to size or shape, cigarette books, and cigarette book covers imported-entered or withdrawn for consumption" measured in pounds, but it is not clear what fraction, if any, of this was rolling papers nor how many rolling papers are produced from a pound of paper. All of these problems with the BATF data are ignored in the calculations below.

Also, some rolling papers were released without tax, but still listed. It is not clear whether these should be treated separately. They are included in the data given here, but the numbers are quite small and so do not appreciably affect the results. Likewise, the BATF also reports the number of cigarette tubes released from factories. These are not included in the estimate of the number of rolling papers because they seem more likely to be used for tobacco cigarettes. At any rate, the number of tubes is also small compared with the number of papers.

An estimate of the number of rolling papers produced is obtained by simply summing the number of books reported times the number of papers in each book. The resulting numbers range from a high of 14.6 billion in FY 1976 to a low of 8.04 billion in FY 1987. These numbers are of the same order of magnitude as Kleiman's estimate of 12 billion marijuana joints per year.

If one ignores the fact that some marijuana is smoked in water pipes and assumes that the average joint contains 0.4 grams of marijuana,[95] these convert to between 3,200 and 4,800 metric tons of marijuana. These quantities are considerably less than those reported by NNICC (See Table 2.9.), but are consistent with estimates used by Kleiman[96] (4,700 metric tons in 1986) and Reuter[97] (4,200 metric tons in 1982).

As Figure 2.12[98] shows, the trend in rolling paper consumption is also consistent with the downward trend in consumption suggested by the High School Senior Survey Data reported in Section 2.5.2.

---

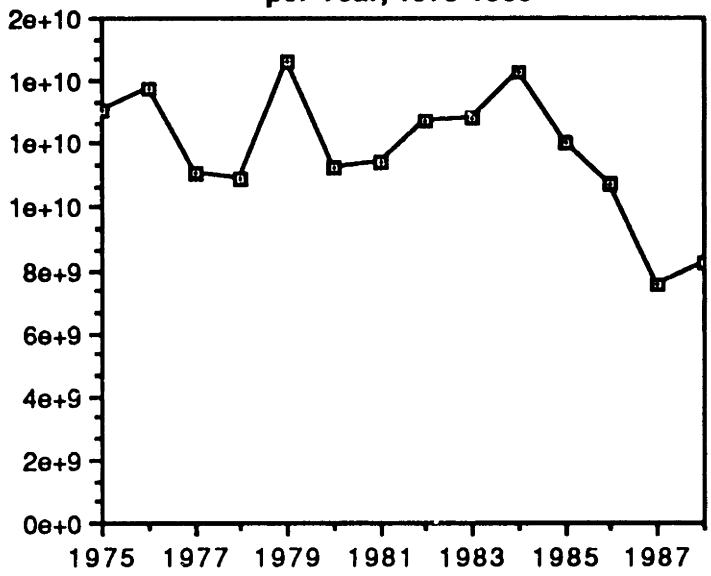
[95]Kleiman mentions this estimate (1989, p.38), but notes that it was made before the increases in potency seen over the last few years.

[96]Kleiman, 1989, p.38.

[97]Reuter, 1984, p.143.

[98]Data from Bureau of Alcohol, Tobacco, and Firearms Summary Statistics on Distilled Spirits, Wine, Beer, Tobacco, Firearms, Enforcement, Taxes, various years.

Figure 2.12
Number of Rolling Papers Consumed per Year, 1975-1988



This suggests that data on rolling papers taxed by the BATF may give some information about marijuana consumption. It also illustrates the kind of indirect evidence that must at times be relied on because the conventional data are so weak.

2.9 Summary

Collecting data on drug markets is difficult, and this is reflected in both the quality and the quantity of data that are available. Many of the existing studies focus on particular subpopulations and generally have more to say about the consumption of illicit drugs than their distribution. Furthermore, existence of mythical numbers (discussed in Section 2.1) that have little if any basis forces one to be cautious about accepting any numbers that do not come directly from a well-run study. Nevertheless, the data can give a feel for the magnitude and nature of the problem, and, as will be seen shortly, some of the data can be used to test theories about how the drug markets operate.

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[15] (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.

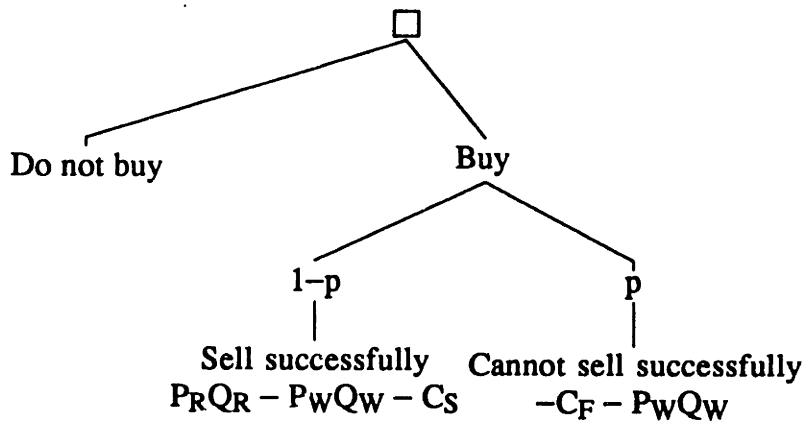
Frequently dealers dilute ("step on") the drugs with adulterants. To avoid confusion, quantities should be understood to refer to the weight of the drugs themselves (excluding adulterants) and all prices are the price per pure unit weight of the drugs. Hence $Q_R \leq Q_W$, and Q_R would only be less 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 $P_R Q_R - P_W Q_W - C_S$ with probability $(1 - p)$ and $-C_F - P_W Q_W$ 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 drug . 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.

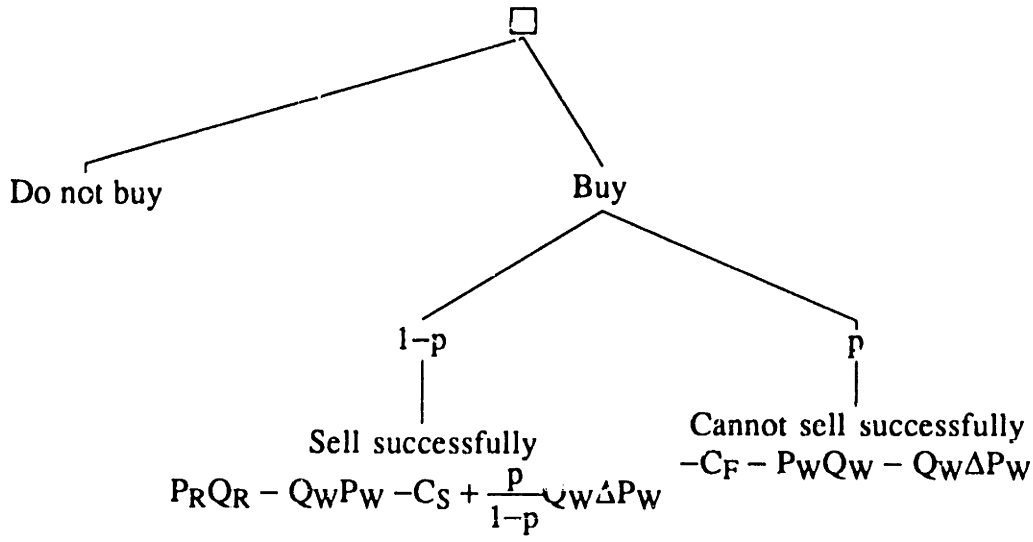


Figure 3.2: Revised Decision Tree With $P_R' = P_R + \frac{Q_W}{Q_R}\left(\frac{1}{1-p} + rT\right)\Delta P_W$

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 $-C_F - P_WQ_W$ and $P_RQ_R - P_WQ_W - C_S$, 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 ($Q'_R = Q'_W\left(\frac{Q_R}{Q_W}\right) = Q_R\left(\frac{P_W}{P'_W}\right)$) and the new retail price is proportionately higher ($P'_R = P_R\left(\frac{P'_W}{P_W}\right)$), 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.

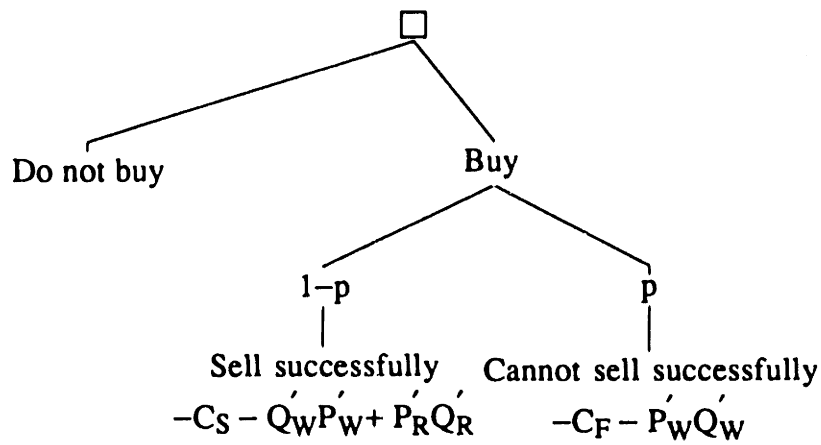


Figure 3.3: Revised Decision Tree for Value-Preserving Model

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

$$Q_R' = \frac{Q_I'}{Q_I} Q_R \tag{3.4}$$

at price

$$P_R' = \frac{P_I'}{P_I} P_R. \tag{3.5}$$

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 $f: R^2 \dashrightarrow R^2$ 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.)

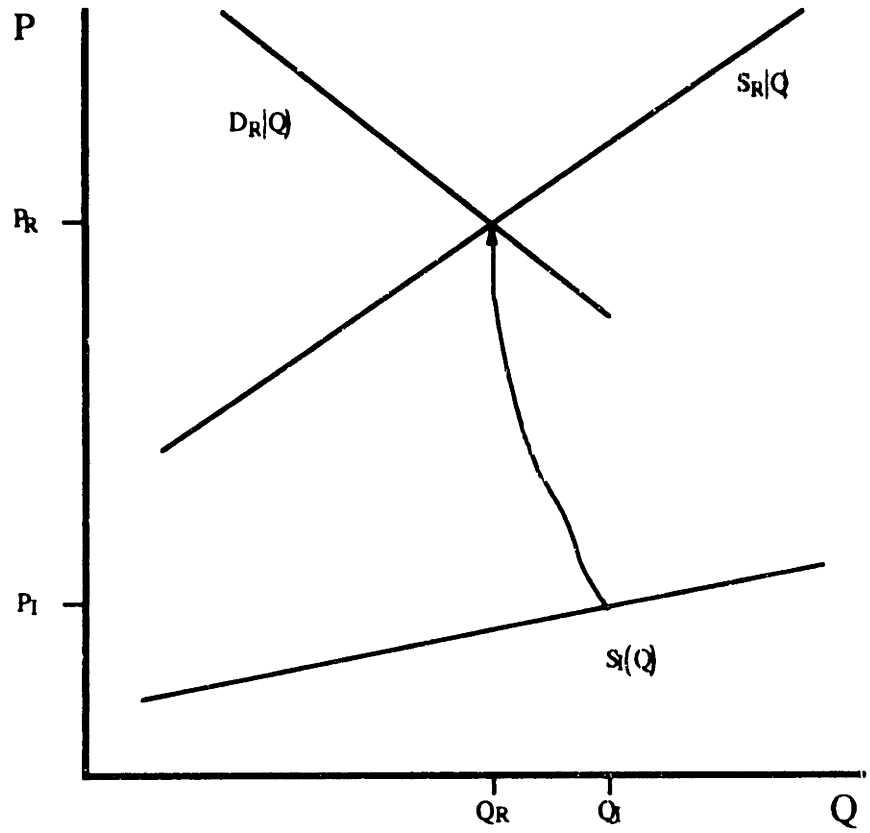


Figure 3.4: Relation Between Import Supply, Retail Supply, and Retail Demand Curves

$S_R(Q)$ 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 **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.

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

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

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

$$Q_R = \frac{b_D - b_R}{a_R - a_D}. \tag{3.14}$$

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

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

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

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

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

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,

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

so for the first interdiction model

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

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

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

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

$$P_R' \qquad P_R + \frac{|a_D|\frac{P_R}{P_I}}{\left(1+\frac{\Delta P_I}{P_I}\right)a_R - a_D}\Delta P_I^* \qquad P_R + \frac{|a_D|\,\alpha}{(1+\alpha)a_R - a_D}P_R$$

$$Q_R' \qquad Q_R - \frac{\frac{P_R}{P_I}}{\left(1+\frac{\Delta P_I}{P_I}\right)a_R - a_D}\Delta P_I^* \qquad Q_R - \frac{\alpha}{(1+\alpha)a_R - a_D}P_R$$

$$P_I' \qquad P_I + \Delta P_I - \frac{a_I\frac{P_R}{P_I}}{\beta\left(\left(1+\frac{\Delta P_I}{P_I}\right)a_R - a_D\right)}\Delta P_I^* \qquad (1+\alpha)P_I - \frac{\alpha\, a_I}{\beta((1+\alpha)a_R - a_D)}P_R$$

$$Q_I' \qquad Q_I - \frac{\frac{P_R}{P_I}}{\beta\left(\left(1+\frac{\Delta P_I}{P_I}\right)a_R - a_D\right)}\Delta P_I^* \qquad Q_I - \frac{\alpha}{\beta((1+\alpha)a_R - a_D)}P_R$$

*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 $\Delta\widehat{P_I} \approx \hat{\alpha}\, P_I$. 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

$$P_R' \approx P_R + (1 + rT)\,\Delta\widehat{P_I} \tag{3.28}$$

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

the additive model predicts that

$$P_R' \approx P_R + \kappa\,\Delta\widehat{P_I} \tag{3.30}$$

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

and for the multiplicative model,

$$P_R' \approx P_R + \frac{P_R}{P_I}\,\Delta\widehat{P_I} \tag{3.32}$$

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

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

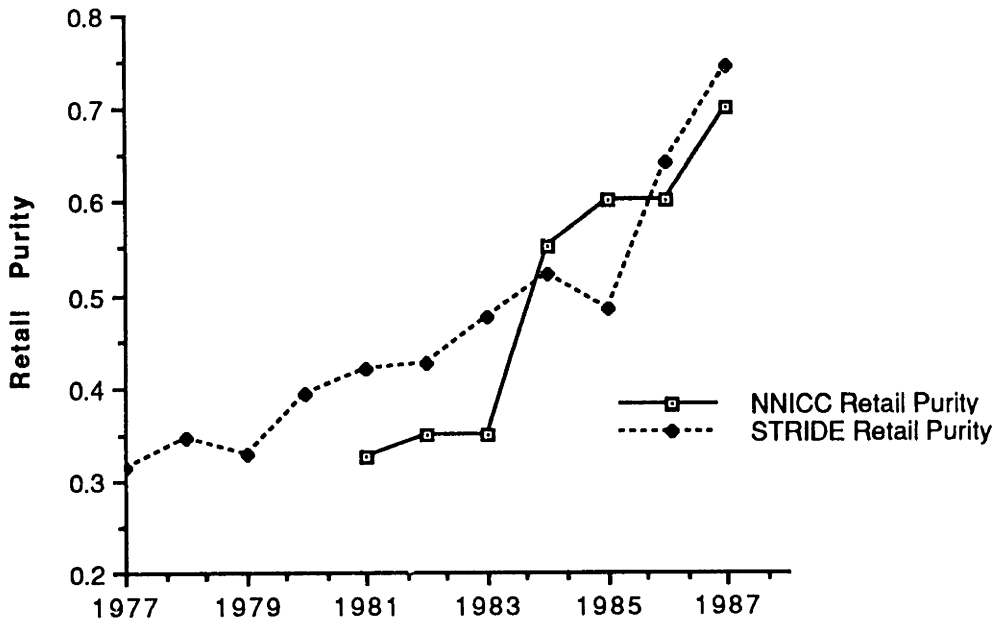


Table 3.5:
Estimates of Retail Cocaine Purity Used to Test the Models

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)

Wholesale (1 kg)	National 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

Retail (1 gm)	National 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)

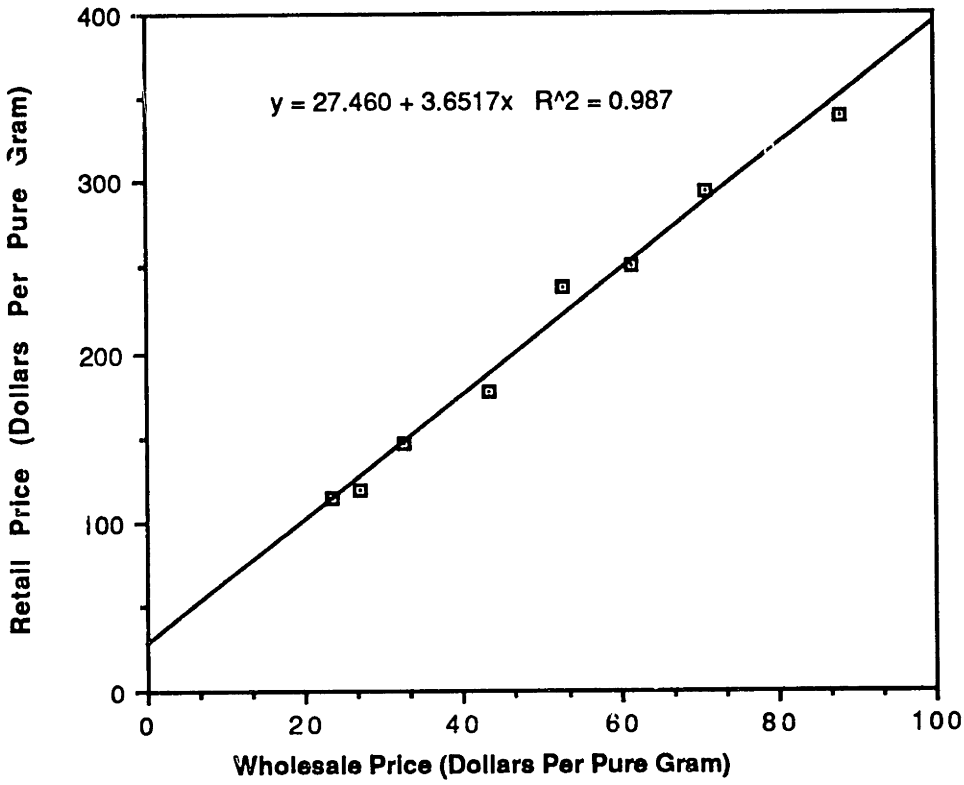


Figure 3.7
Retail vs Wholesale Cocaine Prices
Miami, 1982-1989

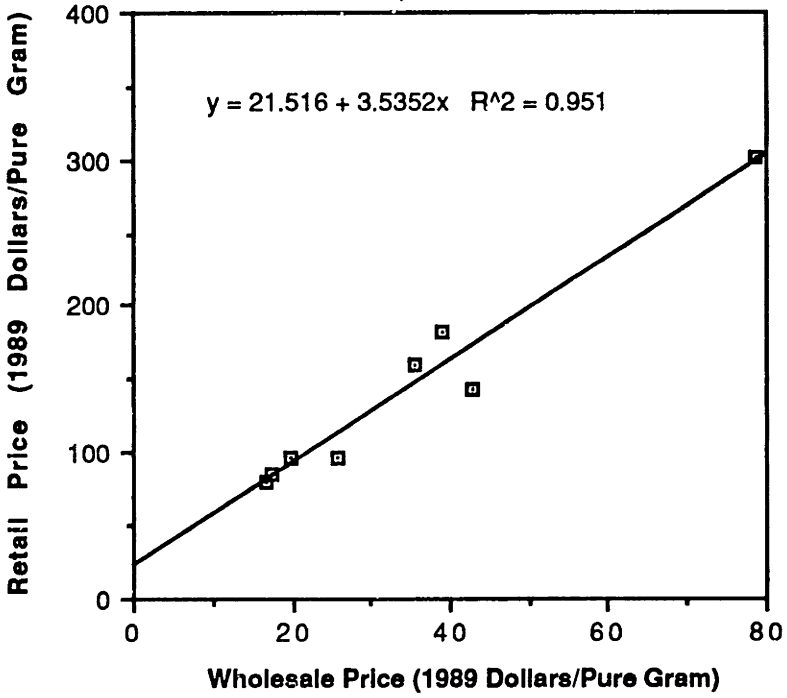


Figure 3.8
Retail vs. Wholesale Cocaine Prices
New York, 1982-1989

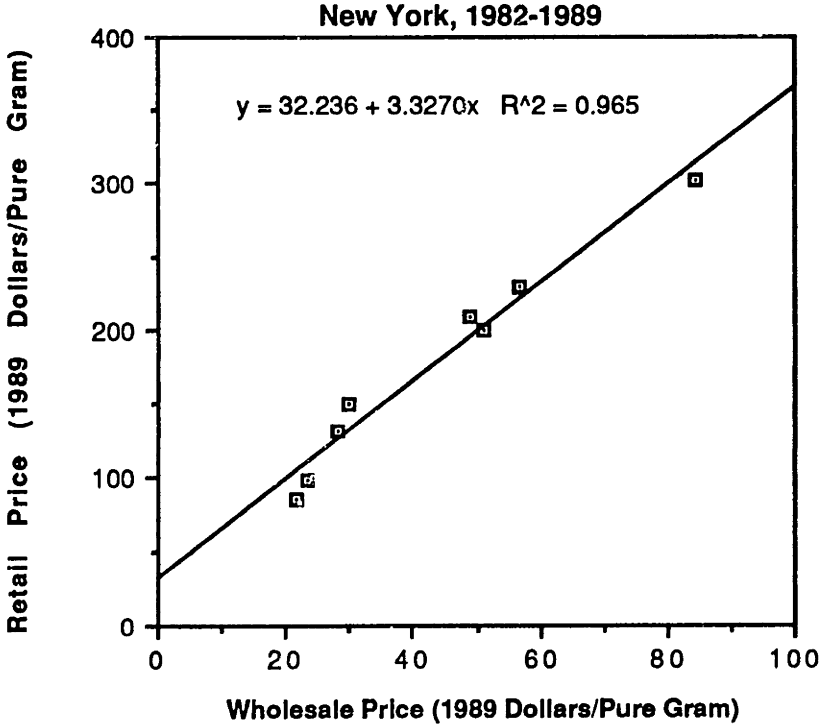


Figure 3.9
Retail vs. Wholesale Cocaine Prices
Chicago, 1982-1989

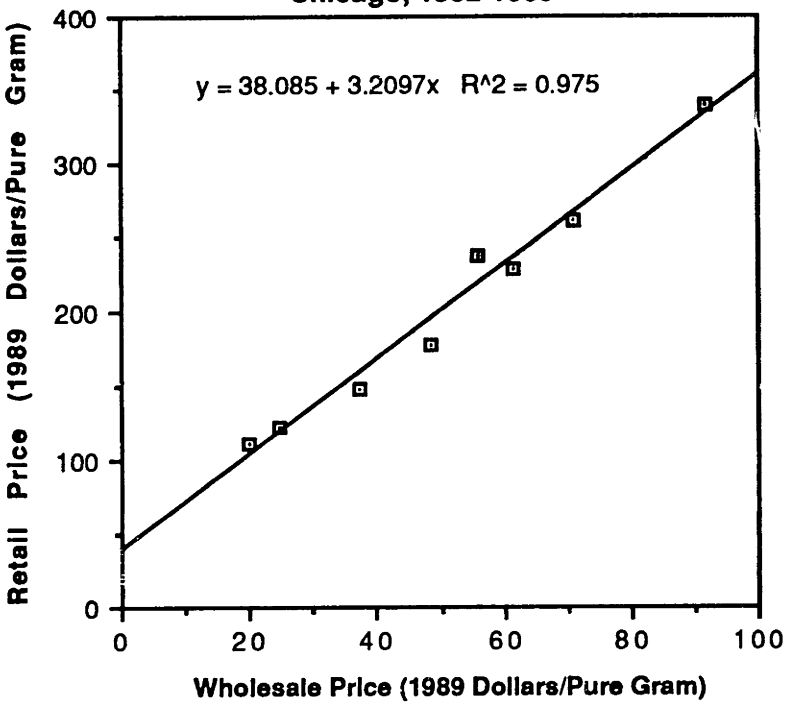
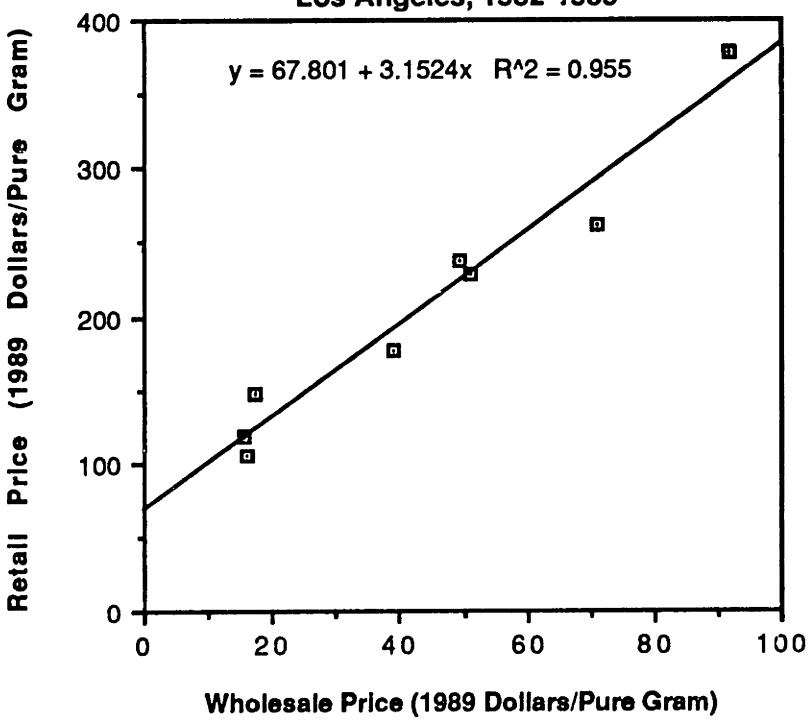


Figure 3.10
Retail vs. Wholesale Cocaine Prices
Los Angeles, 1982-1989



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

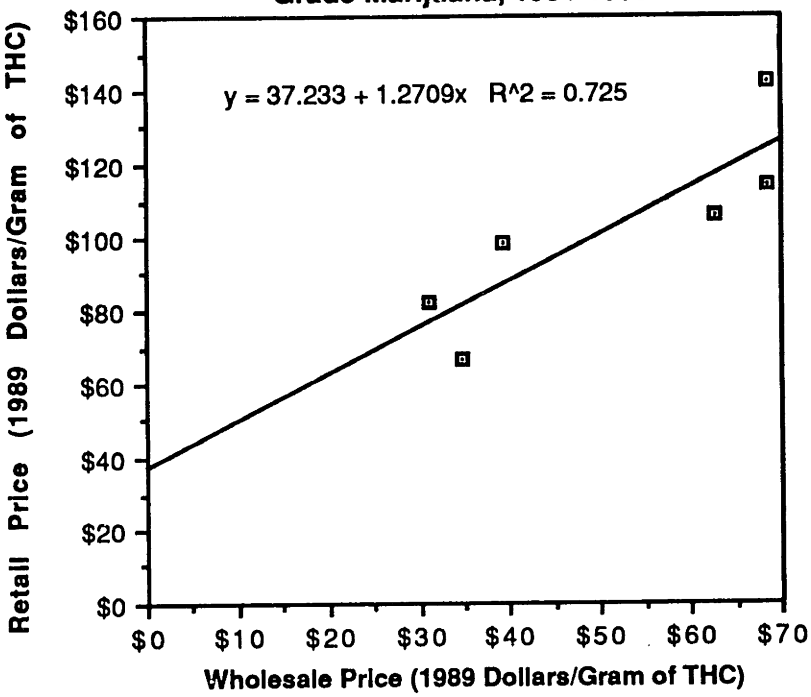
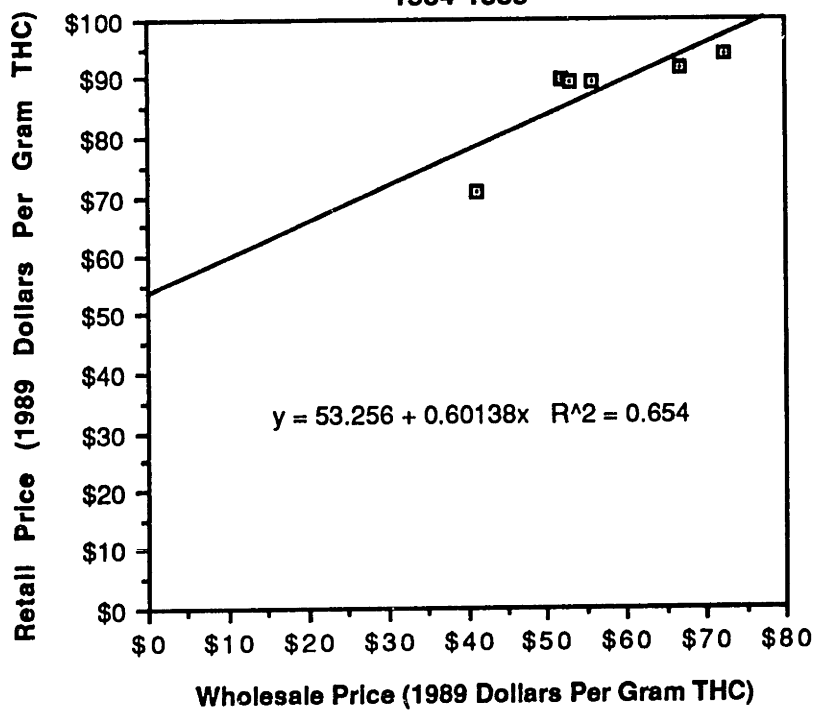


Figure 3.12
Retail vs. Wholesale Prices of Sinsemilla 1984-1989



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

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

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.

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 seli 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

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

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

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

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

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

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.

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

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

	Deal	Do Not Deal
Deal	$\left(\pi q_{max} - \frac{E}{2}, \pi q_{max} - \frac{E}{2}\right)$	$(\pi q_{max} - E, w_0)$
Do Not Deal	$(w_0, \pi q_{max} - E)$	(w_0, w_0)

Suppose

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

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.

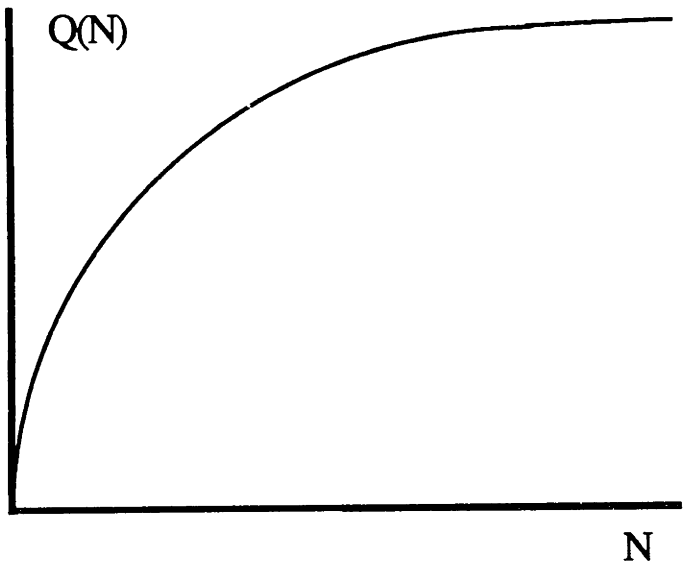
The case when Q(N) is constant yields other important lessons. First, if demand is not responsive to the number of dealers then targeting dealers may not be an effective strategy. Second, if a market is adequately described by this model with Q(N) constant, it is easy to determine whether the crackdown 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

$$Q(N) = \alpha N^{\beta} \qquad \beta \in [0,1] \tag{4.12}$$

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

$$N(E) = \frac{(\pi\alpha)^2 - 2\,w_0\,E + \pi\alpha\sqrt{(\pi\alpha)^2 - 4\,w_0\,E}}{2\,w_0^2}. \tag{4.15}$$

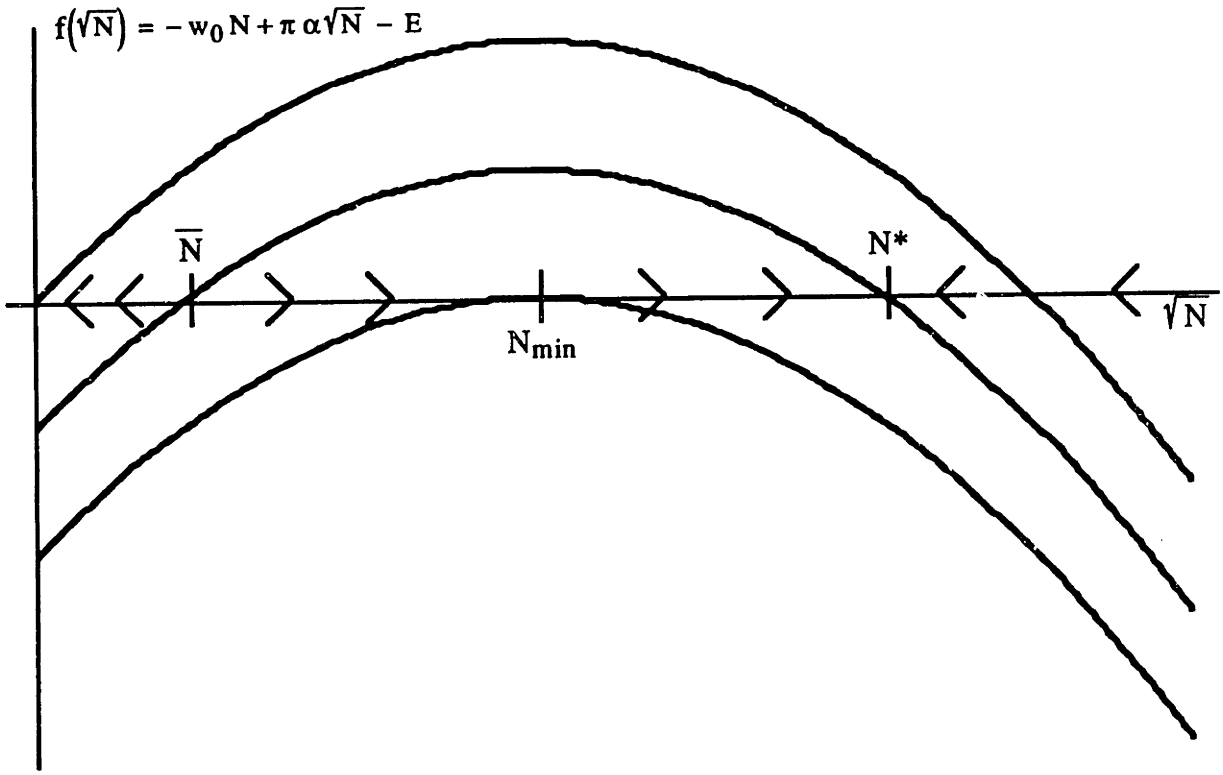
Figure 4.3 shows visually how the number of dealers N(E) is affected by changes in the enforcement pressure. It plots the function

$$f(\sqrt{N}) = -\,w_0\,N + \pi\alpha\sqrt{N} - E \tag{4.16}$$

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

$$n \equiv \frac{N}{N_{max}}, \; q \equiv \frac{Q}{Q_{max}}, \text{ and } e_N \equiv \frac{E}{E_{max}}. \tag{4.25}$$

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".

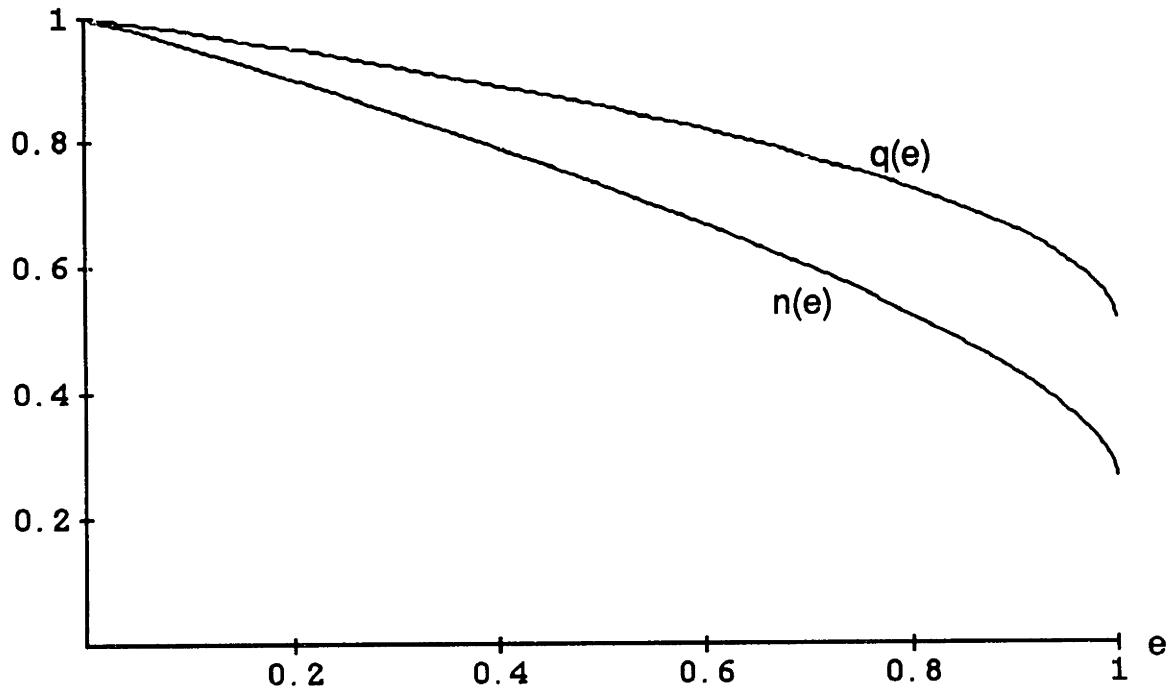
With these definitions one can show that

$$n = \left(\frac{1 + \sqrt{1 - e_N}}{2}\right)^2 \text{ and} \tag{4.26a}$$

$$q = \frac{1 + \sqrt{1 - e_N}}{2}. \tag{4.26b}$$

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
($\beta = 1/2$ and $\gamma = 1$)



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

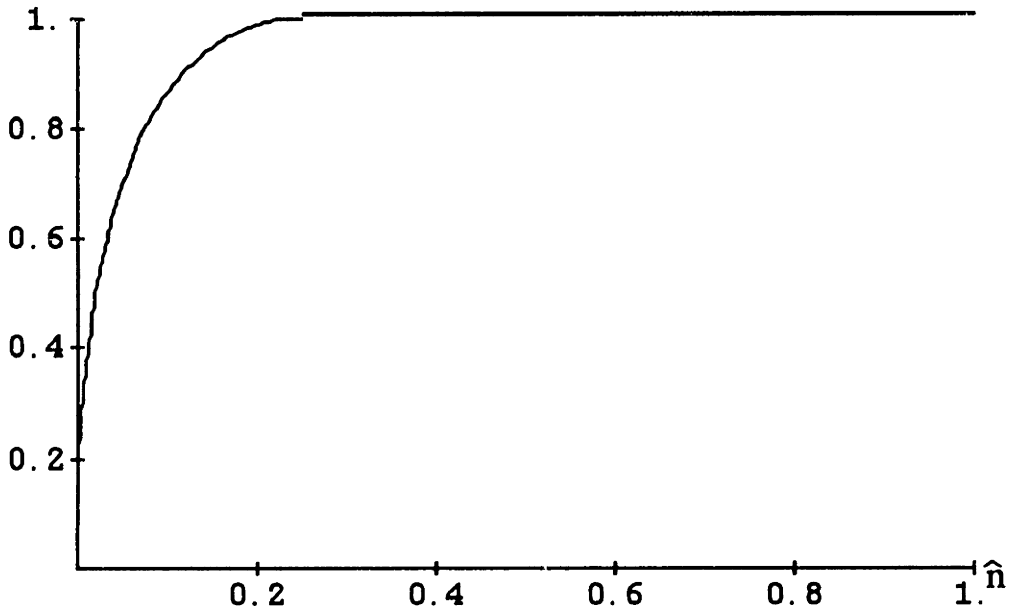
$$\hat{N} \equiv \hat{n}\, N_{max} \tag{4.27}$$

dealers try to start dealing, one needs to maintain a level or pressure equal to

$$\hat{E} = \begin{cases} 4\left(\sqrt{\hat{n}} - \hat{n}\right) E_{max} & \text{if } \hat{n} \leq \frac{1}{4} \\ E_{max} & \hat{n} \geq \frac{1}{4} \end{cases} \tag{4.28}$$

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
($\beta = 1/2$ and $\gamma = 1$)



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

$$\frac{dN}{dt} = c_1\left[\frac{\pi \alpha N^{\beta}}{N} - \left(\frac{E}{N}\right)^{\gamma} - w_0\right] = F(N). \qquad (4.29)$$

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

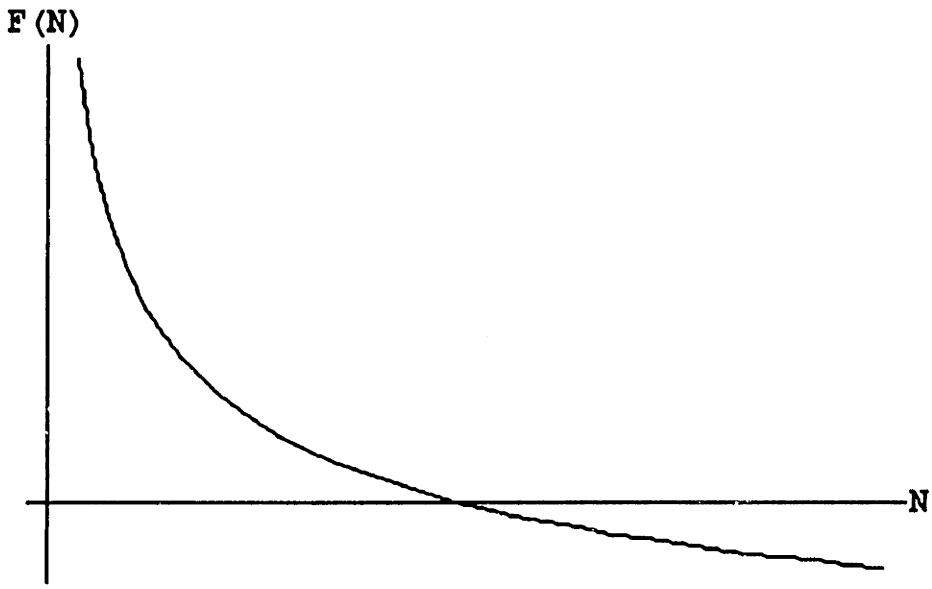
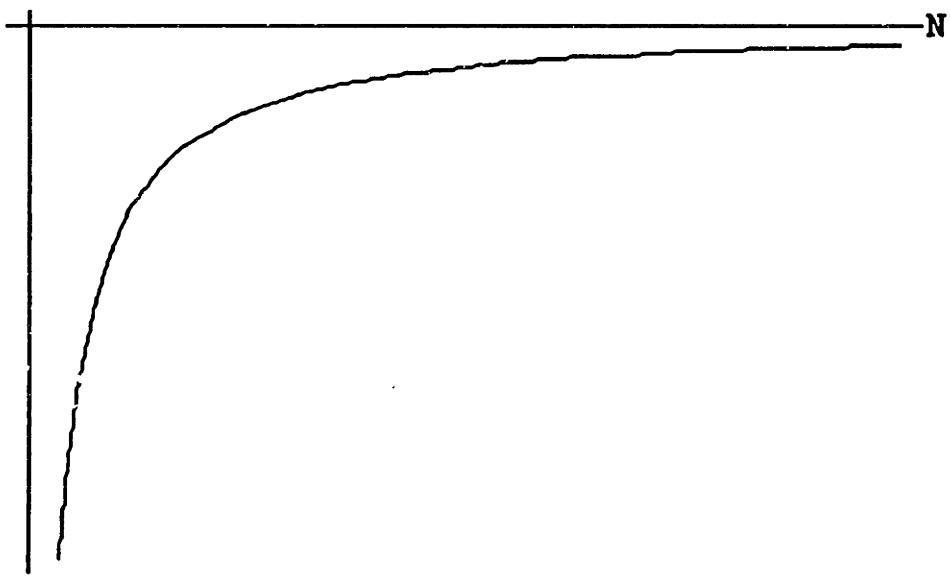


Figure 4.7:

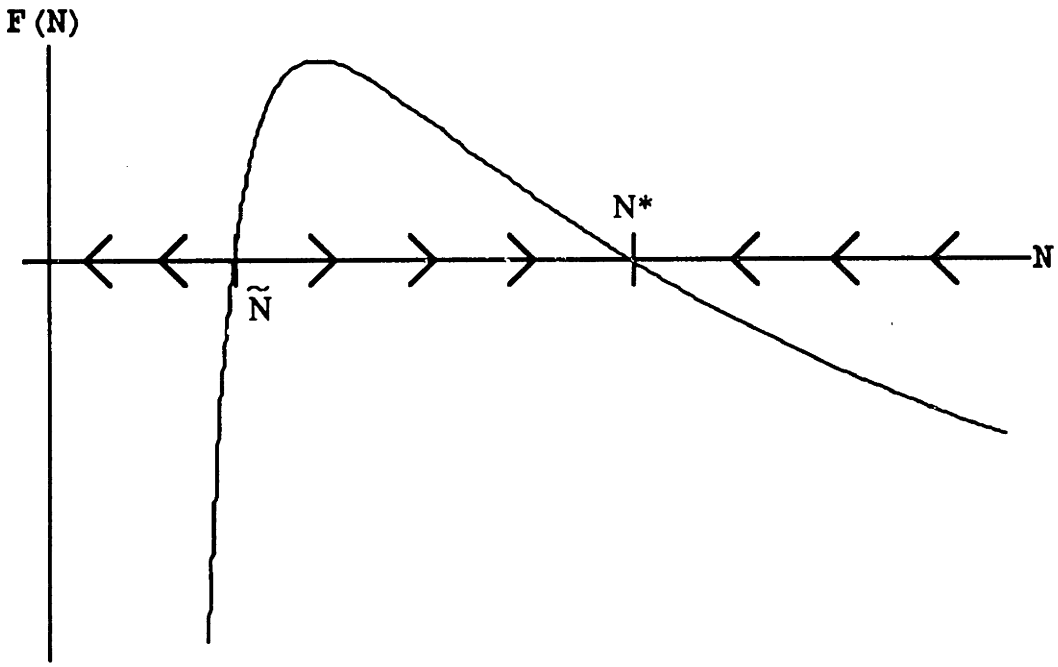
Enforcement's Contribution to $\frac{dN}{dt}$



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 $N \approx \bar{N}$. 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

$$E_{max} = \left(\frac{1-\beta}{\gamma - 1 + \beta}\right)^{1/\gamma} \left(\frac{\gamma - 1 + \beta}{\gamma}\right)^{1/(1-\beta)} w_0^{1/\gamma} N_{max}. \quad (4.37)$$

This reduces to

$$= (1-\beta)\beta^{\beta/(1-\beta)} w_0 N_{max} \quad (4.38)$$

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

$$E_{max} = \left(\frac{1-\beta}{\gamma - 1 + \beta}\right)^{1/\gamma} \left(\frac{\gamma - 1 + \beta}{\gamma}\right)^{1/(1-\beta)} w_0^{(1-\gamma)/\gamma} \pi Q_{max} \quad (4.39)$$

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.

---

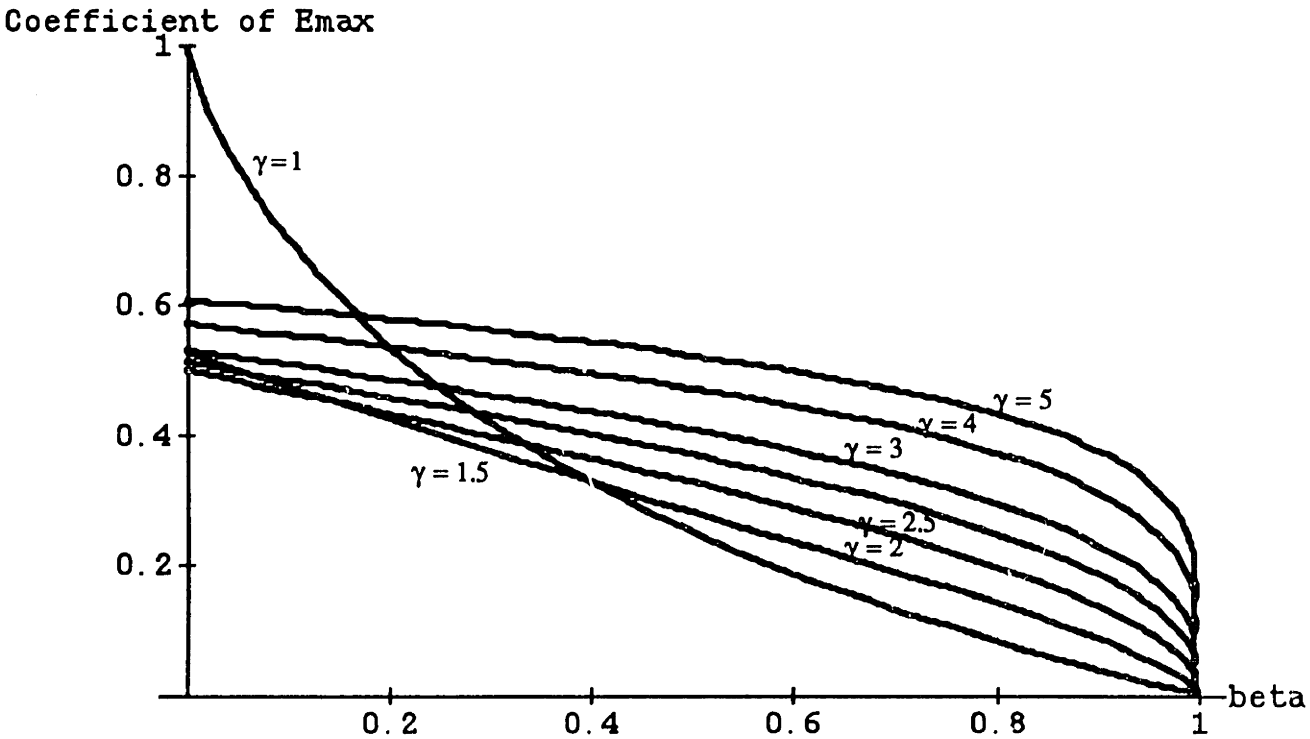
[57]Kleiman, 1988a, pp.25-26 and p.29.

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

$$C(\beta,\gamma) \equiv \left(\frac{1-\beta}{\gamma - 1+\beta}\right)^{1/\gamma}\left(\frac{\gamma - 1+\beta}{\gamma}\right)^{1/(1-\beta)}. \qquad (4.40)$$

Figure 4.9 plots this expression as a function of β for various values

Figure 4.9:
Coefficient of E_{max}, $C(\beta,\gamma)$

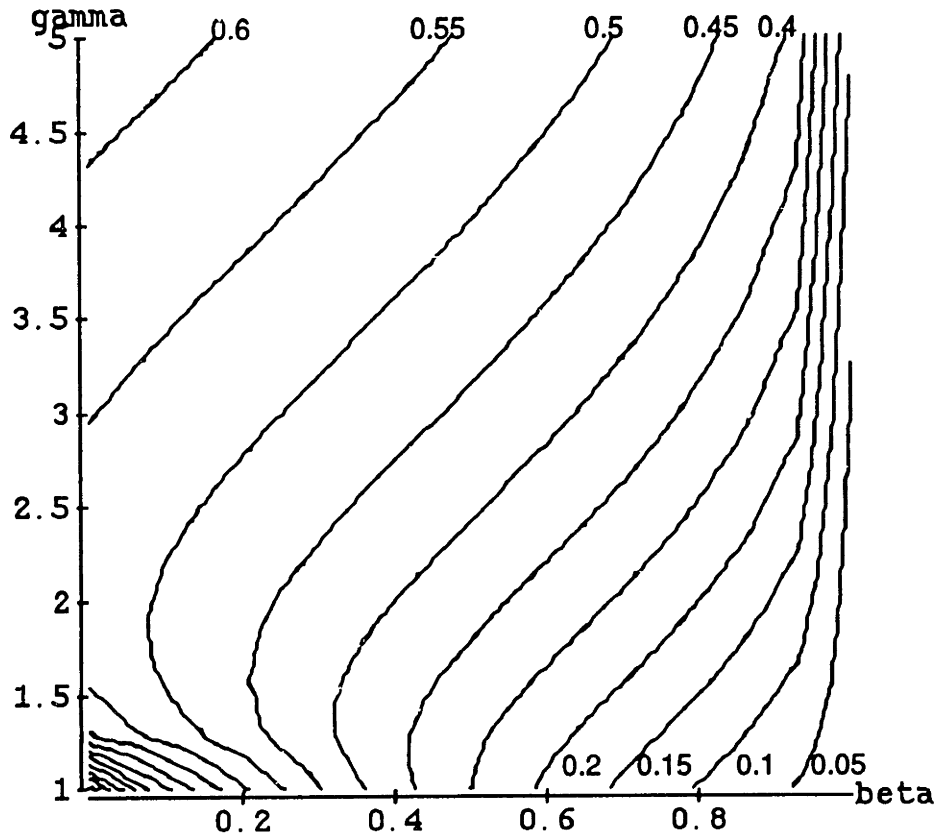


of γ. These plots show how the ratio of enforcement to market size needed to collapse the market might vary from market to market.

Figure 4.10 displays the same information in a different way; it plots the level curves of $C(\beta,\gamma)$ on the β-γ plane.

The figures show that $C(\beta,\gamma)$, 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 $C(\beta,\gamma)$ decreases in β, it would probably be easier, perhaps even much easier, to collapse the second market.

Figure 4.10:
Level Curves of $C(\beta,\gamma)$
(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

$$N_{min} = \left(1 - \frac{1-\beta}{\gamma}\right)^{1/(1-\beta)} \left(\frac{\pi\alpha}{w_0}\right)^{1/(1-\beta)} \tag{4.41}$$

$$= \left(1 - \frac{1-\beta}{\gamma}\right)^{1/(1-\beta)} N_{max}.$$

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

$$Q_{min} = \alpha N_{min}^{\beta} = \alpha\left(1 - \frac{1-\beta}{\gamma}\right)^{\beta/(1-\beta)} N_{max}^{\beta}$$

$$= \left(1 - \frac{1-\beta}{\gamma}\right)^{\beta/(1-\beta)} Q_{max}. \tag{4.42}$$

So Q_{min} is proportional to Q_{max}.

The proportionality constants differ. In particular, defining

$$n_{min} \equiv \frac{N_{min}}{N_{max}} = \left(1 - \frac{1-\beta}{\gamma}\right)^{1/(1-\beta)} \text{ and} \tag{4.43}$$

$$q_{min} \equiv \frac{Q_{min}}{Q_{max}} = \left(1 - \frac{1-\beta}{\gamma}\right)^{\beta/(1-\beta)}, \tag{4.44}$$

one can see that

$$q_{min} = n_{min}^{\beta}. \tag{4.45}$$

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

$$\frac{Q}{N} = \frac{Q_{min}}{N_{min}} = \left(\frac{\gamma}{\gamma + \beta - 1}\right)\frac{w_0}{\pi} > \frac{w_0}{\pi}. \tag{4.46}$$

Fig 4.11:
n_{min} as a Function of β for Various γ

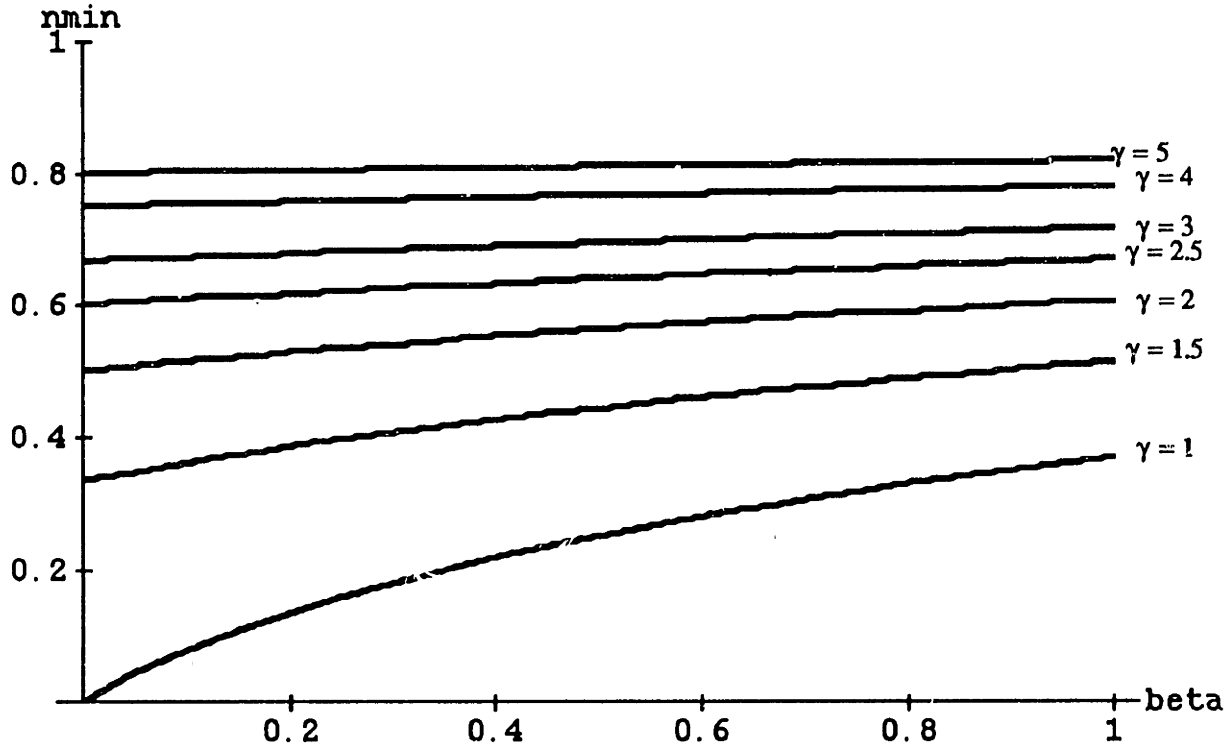
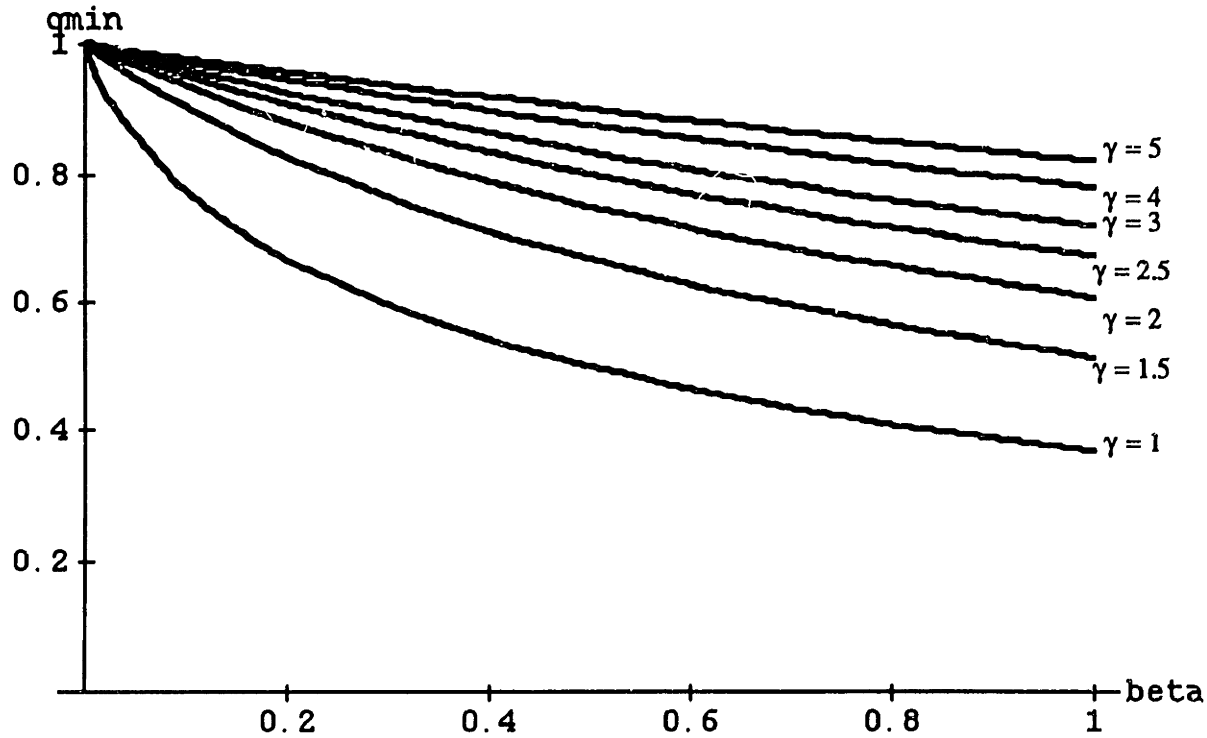


Fig 4.12:
q_{min} as a Function of β for Various γ



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 plots e(n) for β = 0.1, 0.3, 0.5, 0.7, and 0.9 for γ = 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 γ = 1.0 and β = 0.1, 0.3, 0.5, 0.7, and 0.9

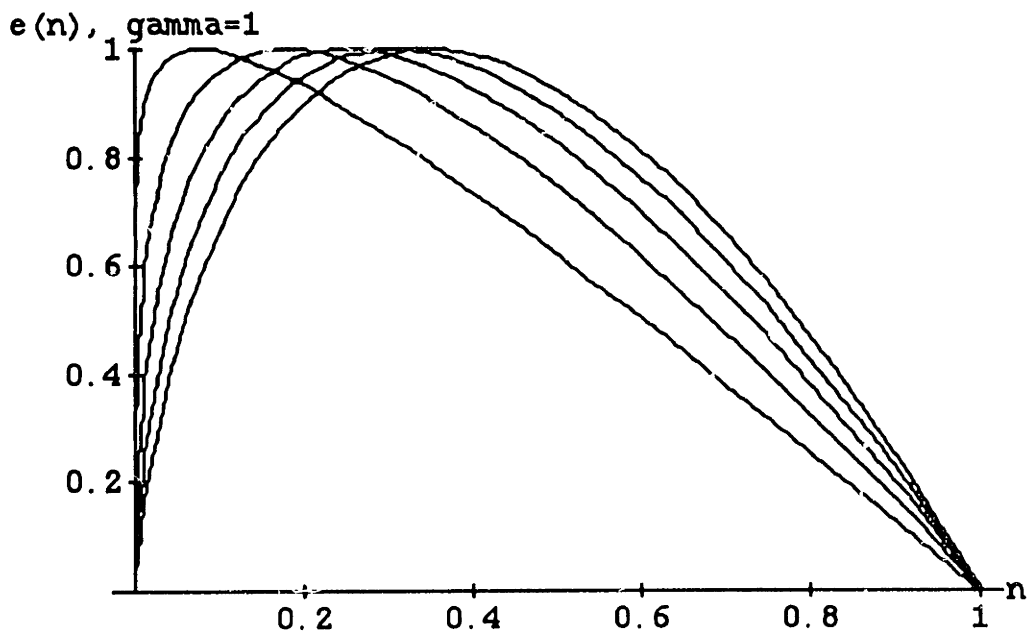


Figure 4.13b:
$e_N(n)$ for $\gamma = 1.5$ and $\beta = 0.1, 0.3, 0.5, 0.7,$ and 0.9

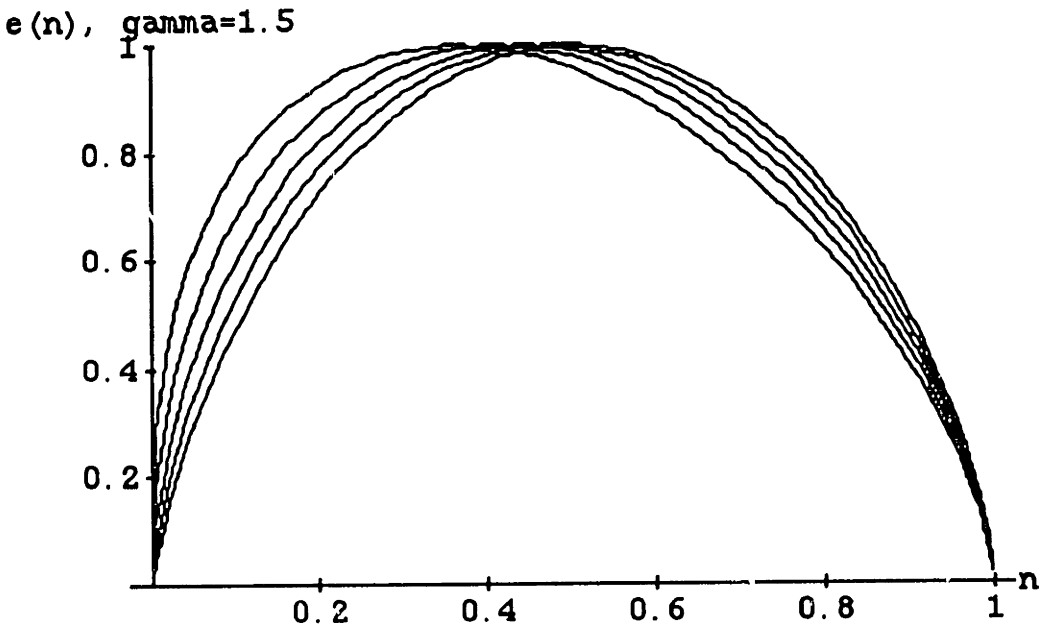


Figure 4.13c:
$e_N(n)$ for $\gamma = 2.0$ and $\beta = 0.1, 0.3, 0.5, 0.7,$ and 0.9

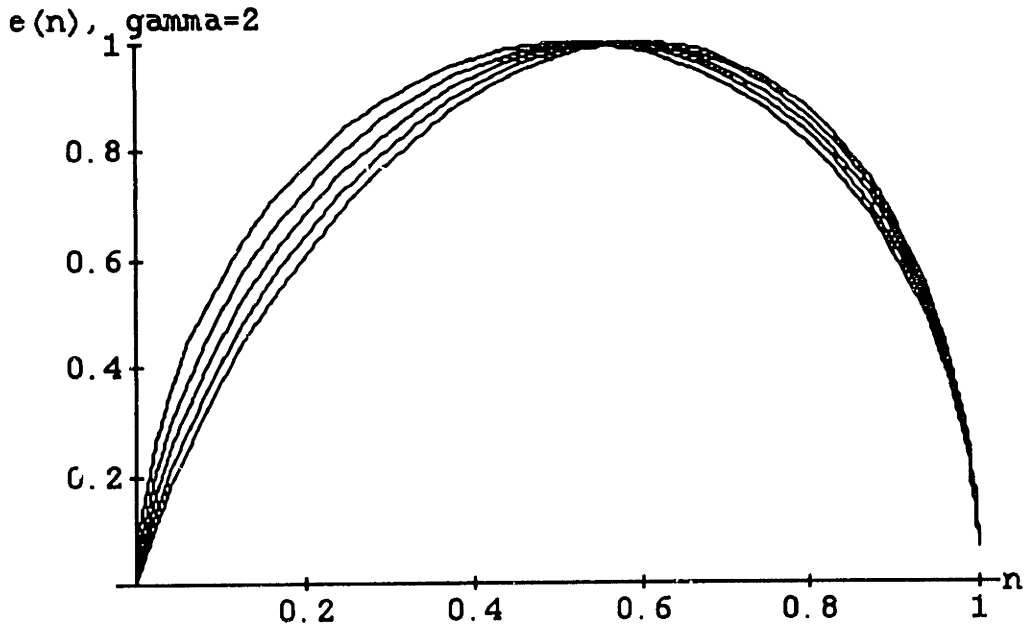


Figure 4.13d:
$e_N(n)$ for $\gamma = 2.5$ and $\beta = 0.1, 0.3, 0.5, 0.7,$ and 0.9

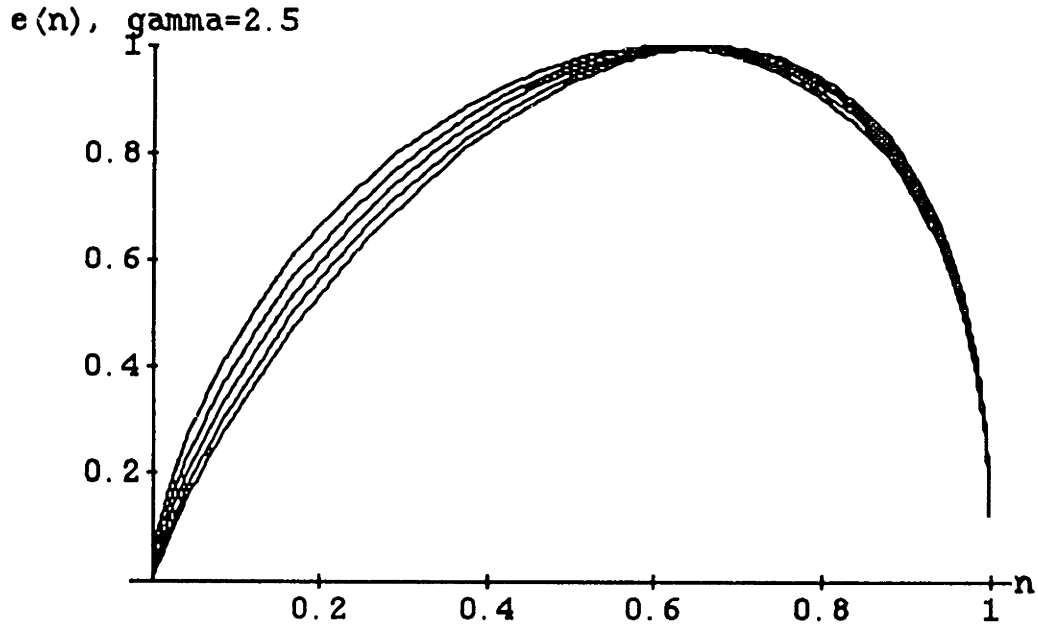
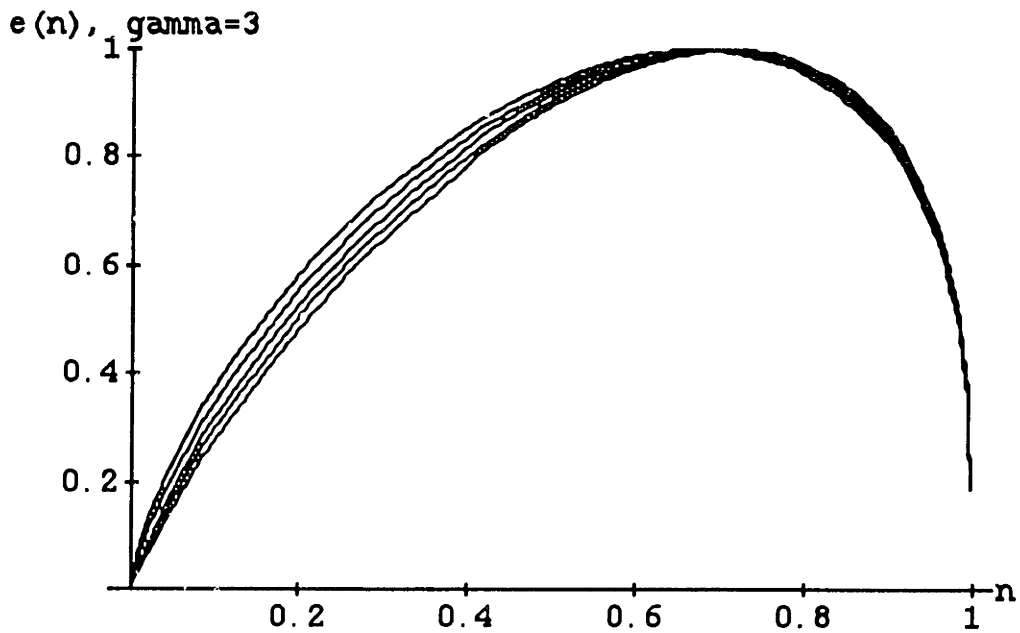


Figure 4.13e:
$e_N(n)$ for $\gamma = 3.0$ and $\beta = 0.1, 0.3, 0.5, 0.7,$ and 0.9



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 u_{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

$$\frac{de(n,\beta,\gamma)}{d\beta} < 0, \tag{4.51a}$$

$$Sgn\left(\frac{d^2e(n,\beta,\gamma)}{d\beta^2}\right) = Sgn\left(n^{\beta-1} - \gamma\right), \tag{4.51b}$$

$$Sgn\left(\frac{de(n,\beta,\gamma)}{dn}\right) < 0, \tag{4.51c}$$

$$\frac{d^2e(n,\beta,\gamma)}{dn^2} < 0, \tag{4.51d}$$

$$Sgn\left(\frac{de(n,\beta,\gamma)}{d\gamma}\right) = Sgn\left(n - \left(\frac{1}{2}\right)^{1/(1-\beta)}\right), \text{ and} \tag{4.51e}$$

$$\frac{d^2e(n,\beta,\gamma)}{d\gamma^2} < 0 \quad \text{iff} \quad \left(\frac{1}{2}\right)^{1/(1-\beta)} < n < \left(\frac{1}{1+e^{-2\gamma}}\right)^{1/(1-\beta)}. \tag{4.51f}$$

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

$$\left(\frac{\pi\alpha}{2w_0} + \frac{1}{2}\sqrt{\left(\frac{\pi\alpha}{2w_0}\right)^2 - \frac{4E^{\gamma}}{w_0}}\right)^{2/\gamma} - \left(\frac{\pi\alpha}{2w_0} - \frac{1}{2}\sqrt{\left(\frac{\pi\alpha}{2w_0}\right)^2 - \frac{4E^{\gamma}}{w_0}}\right)^{2/\gamma}$$

$$= N_{max}\left(\left[\frac{1}{2}\left(1 + \sqrt{1 - e_N^{\gamma}}\right)\right]^{2/\gamma} - \left[\frac{1}{2}\left(1 - \sqrt{1 - e_N^{\gamma}}\right)\right]^{2/\gamma}\right). \tag{4.58}$$

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

$$N = \left(\frac{\pi\alpha}{3w_0}\left[1 + 2\,\mathrm{Cos}\left[\frac{1}{3}\,\mathrm{Cos}^{-1}\left(1 - 2e_N^{\gamma}\right)\right]\right]\right)^{3/\gamma}. \tag{4.61}$$

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

$$\widetilde{e_N} = \frac{E}{w_0{}^{1/\gamma}\,N_{max}}$$

$$= \left(\frac{1-\beta}{\gamma+\beta-1}\right)^{1/\gamma}\left(\frac{\gamma+\beta-1}{\gamma}\right)^{1/(1-\beta)} e_N. \qquad (4.66)$$

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

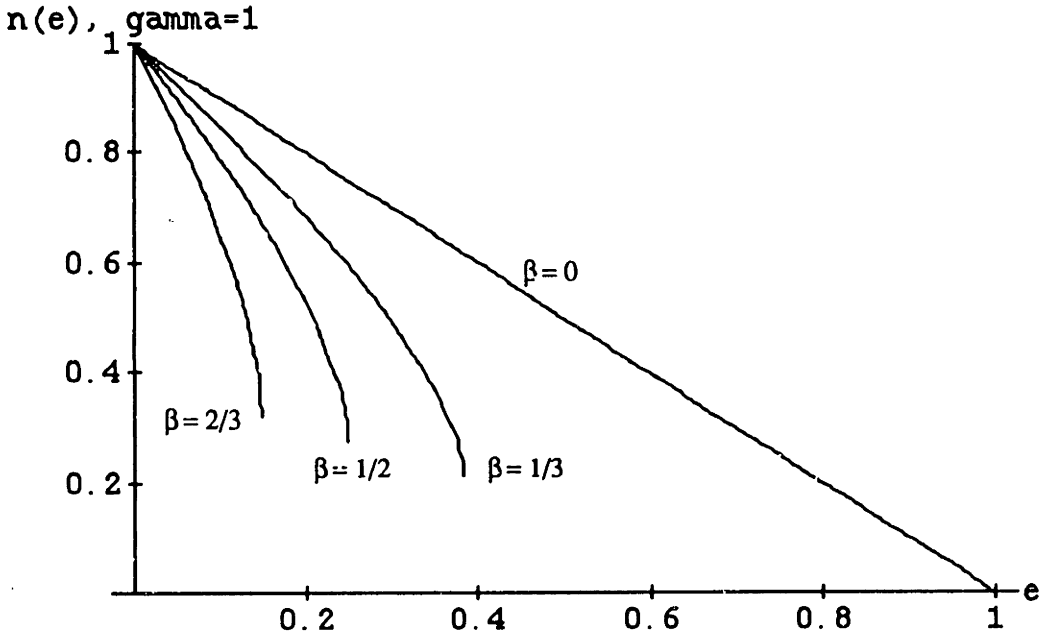
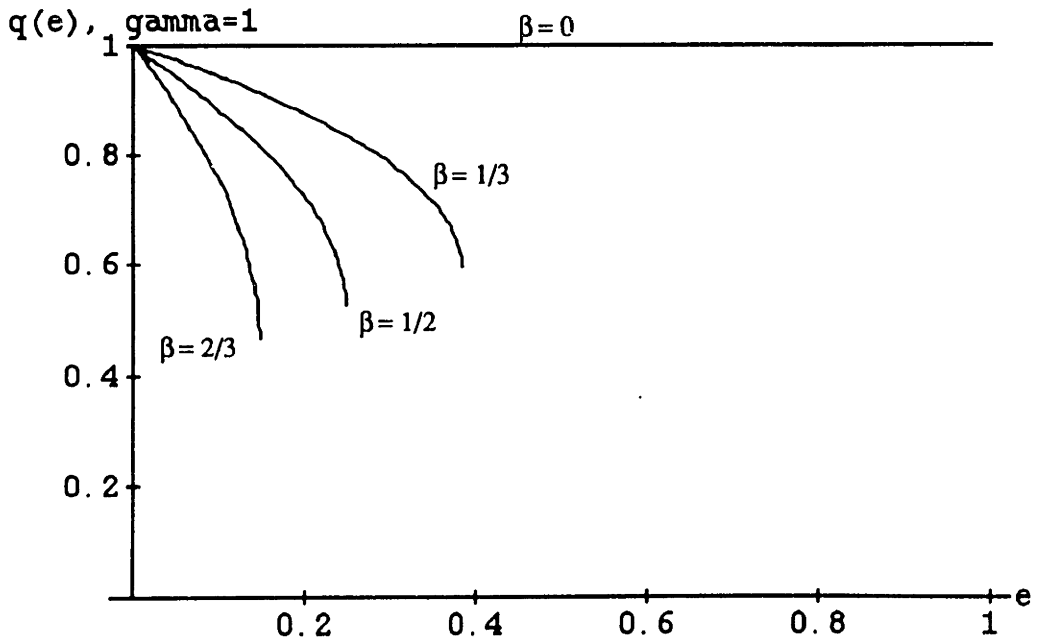


Figure 4.14b:
$Q(\tilde{e}_N)$ for $\gamma = 1$ and $\beta = 0$, 1/3, 1/2 and 2/3



Figures 4.15a,b are the corresponding plots for $\beta = 0, 1/4$, and $1/2$ and $\gamma = 3/2$.

Figure 4.15a:
$N(\tilde{e}_N)$ for $\gamma = 1.5$ and $\beta = 0, 1/4, 1/2$

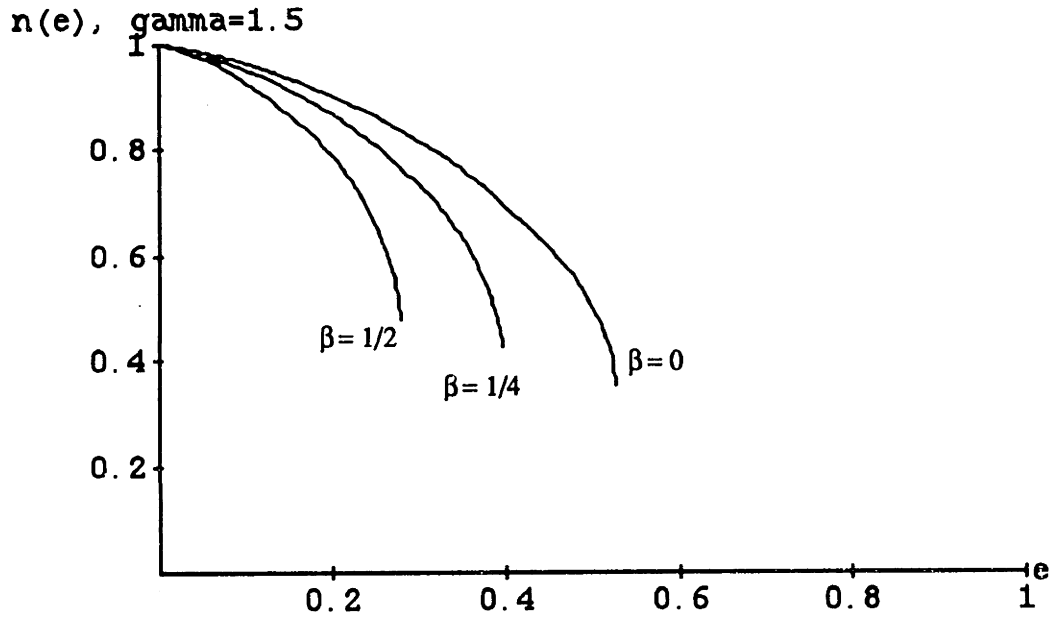
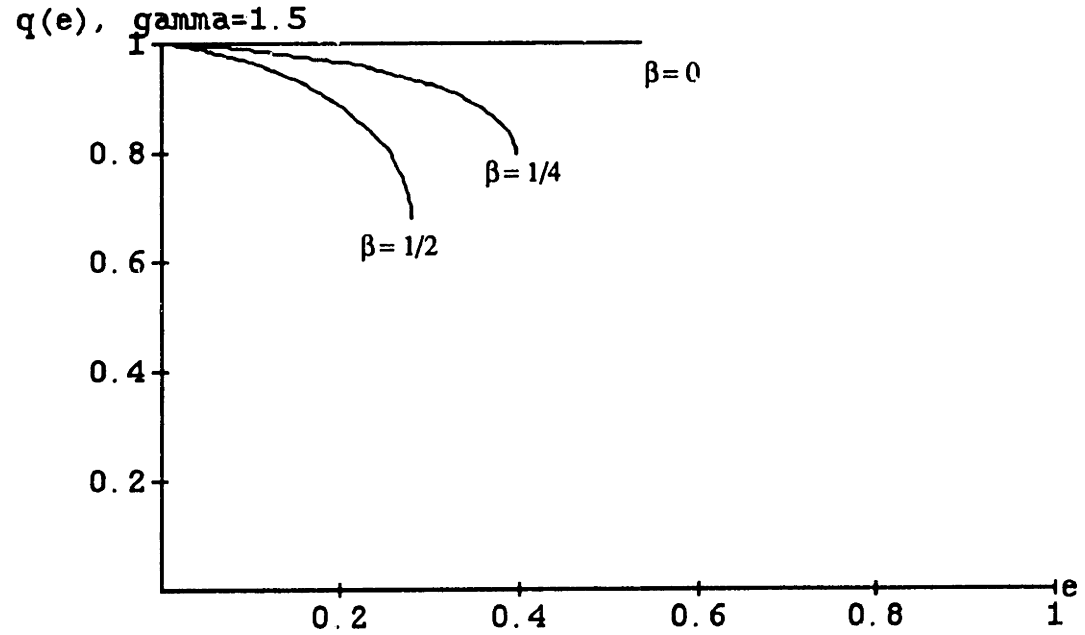


Figure 4.15b:
$Q(\tilde{e}_N)$ for $\gamma = 1.5$ and $\beta = 0, 1/4, 1/2$



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 $\beta = 0, 1/3$

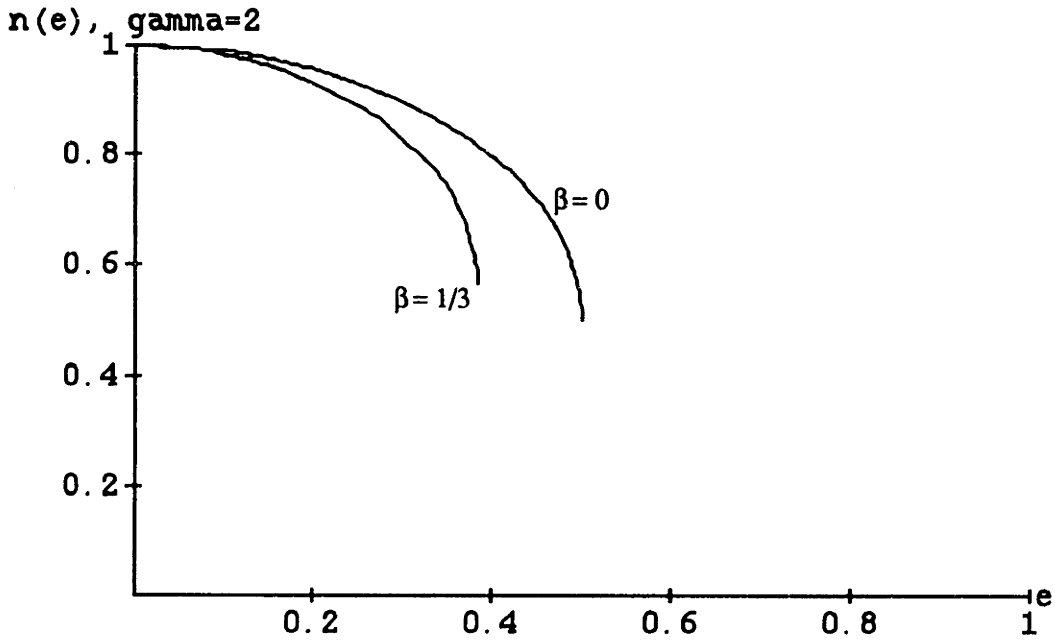
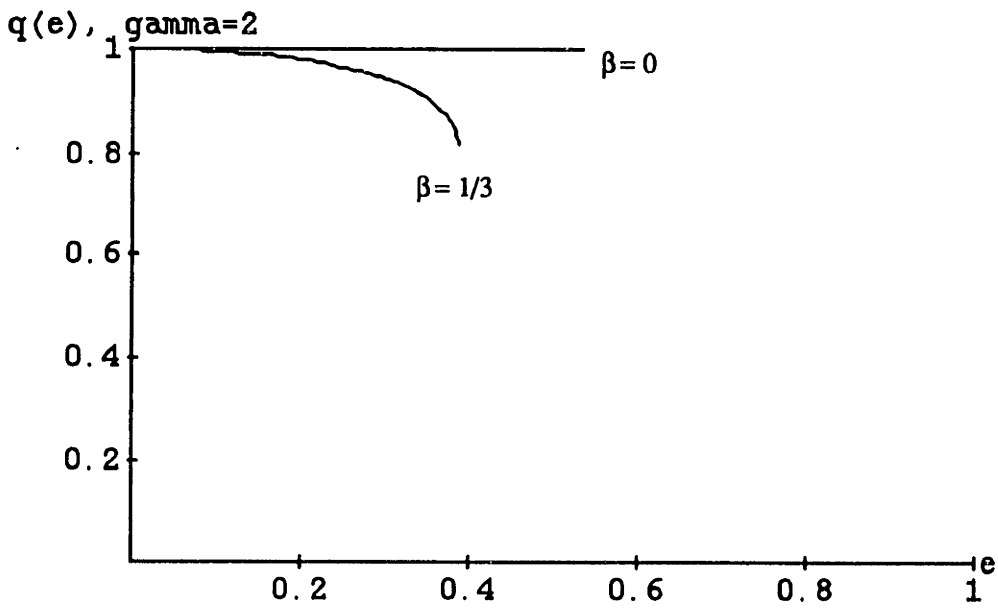


Figure 4.16b:
$Q(\tilde{e}_N)$ for $\gamma = 2.0$ and $\beta = 0, 1/3$



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

$$\frac{\pi \alpha \hat{N}^{\beta}}{\hat{N}} - \left(\frac{E}{\hat{N}}\right)^{\gamma} - w_0 < 0, \tag{4.67}$$

i.e. that

$$E > \left[\pi \alpha \hat{N}^{\beta-1} - w_0\right]^{1/\gamma} \hat{N} \tag{4.68}$$

$$= \left[\hat{n}^{\gamma+\beta-1} - \hat{n}^{\gamma}\right]^{1/\gamma} w_0^{1/\gamma} N_{max},$$

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,

$$E_{max}{}^{3} = \text{Max}\{ E_{max}{}^{1}, \alpha E_{max}{}^{2} \}. \qquad (4.71)$$

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

$$\frac{\%\Delta N_{max}}{\%\Delta \alpha} = \frac{d N_{max}}{d \alpha} \frac{\alpha}{N_{max}} = \frac{1}{1 - \beta} \tag{4.72}$$

and

$$\frac{\%\Delta Q}{\%\Delta N} = \frac{dQ}{dN} \frac{N}{Q} = \beta. \tag{4.73}$$

Note, the first result is the same whether one measures market size in terms of dealers or number of sales because

$$\frac{\%\Delta Q_{max}}{\%\Delta \alpha} = \frac{d Q_{max}}{d \alpha} \frac{\alpha}{Q_{max}} = \frac{1}{1 - \beta}. \tag{4.74}$$

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

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,

$$E_{max}{}^{3} = \text{Max}\{ E_{max}{}^{1}, \alpha E_{max}{}^{2} \}. \qquad (4.71)$$

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 $(1 - f)\,\alpha$:

$$\min_{0 \le f \le 1} z(f) = E_D(f) + \left(\frac{1-\beta}{\gamma+\beta-1}\right)^{1/\gamma}\left(\frac{\gamma+\beta-1}{\gamma}\right)^{1/(1-\beta)} w_0^{1/\gamma}\, N_{max}$$

$$= \min_{0 \le f \le 1} z(f) = E_D(f) + \left(\frac{1-\beta}{\gamma+\beta-1}\right)^{1/\gamma}\left(\frac{\gamma+\beta-1}{\gamma}\right)^{1/(1-\beta)} w_0^{1/\gamma}\left(\frac{\pi\,(1-f)\,\alpha}{w_0}\right)^{1/(1-\beta)}. \tag{4.75}$$

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

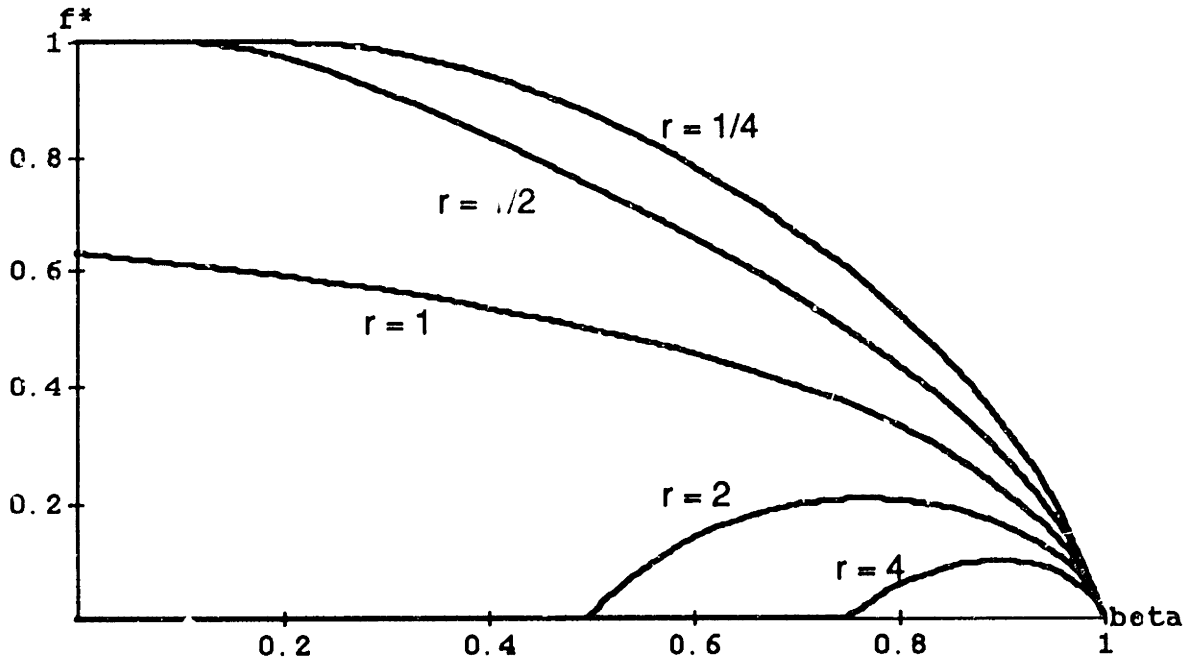
$$r \equiv \frac{c\,Q_{max}}{E_{max}}. \tag{4.77}$$

Then Equation 4.76 can be rewritten as

$$f^* = \text{MAX}\left\{1 - \left((1-\beta)\, r\right)^{\frac{1-\beta}{\beta}},\ 0\right\}$$

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

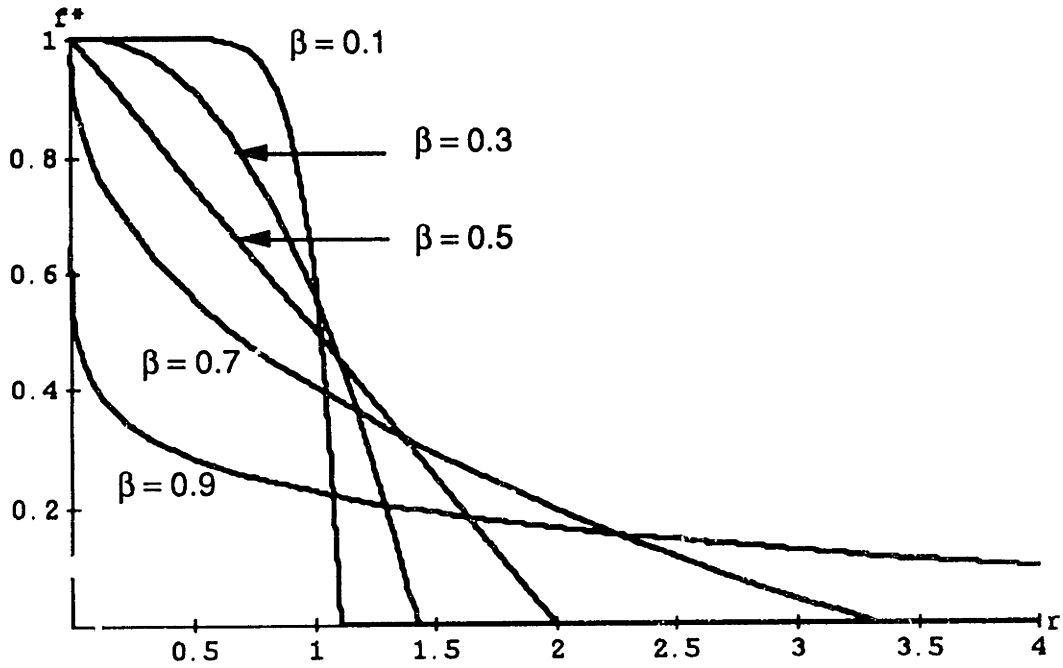


Figure 4.18 plots f* as a function of r for various values of β. It too shows something interesting. Again if β is large, f* is relatively small. Enforcement against dealers works best when β is close to one and dealers are in short supply. But for a wide range of r, some mixture of demand reduction 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.

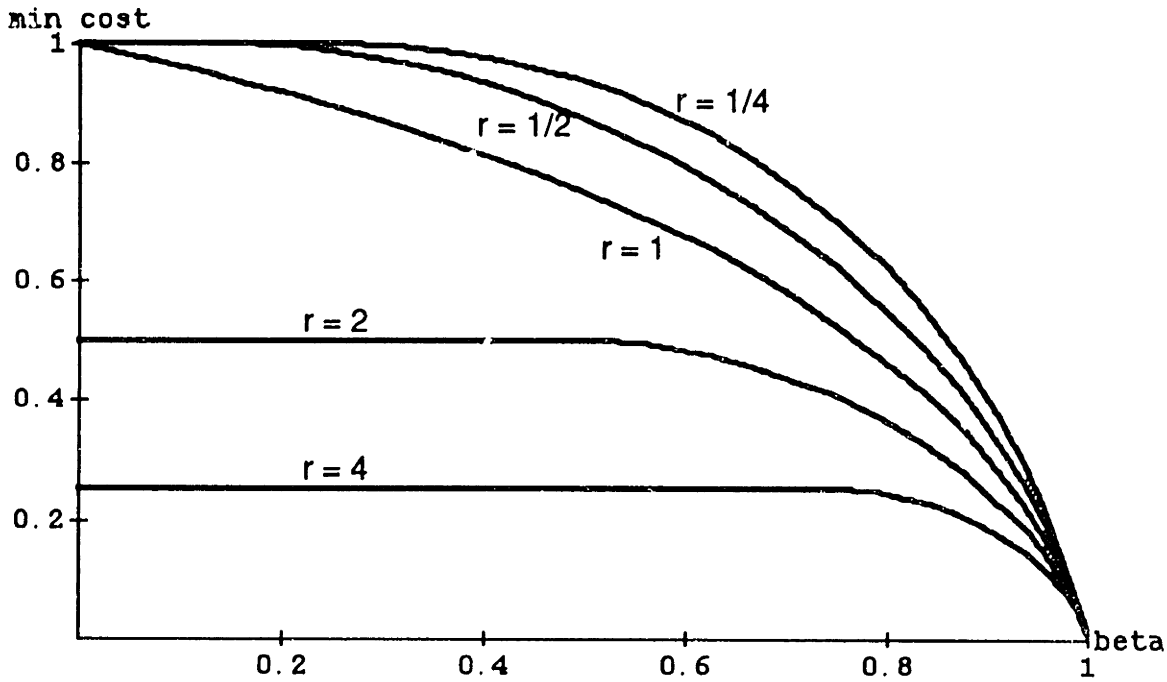
The minimum cost of eliminating the market is

$$c\,Q_{max}\,f^* + E_{max}\,(1 - f^*)^{1/(1-\beta)}$$

$$= c\,Q_{max}\left(1 - \beta\left((1-\beta)\,r\right)^{(1-\beta)/\beta}\right). \qquad (4.78)$$

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:

$$\min_{0 \le f \le 1} Q(\alpha(f), E(f)) \tag{4.79}$$

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	$w = w0$
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	$f N_0$	$w = w0$
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	$w0$
Final equilibrium				
∞	$(1 + f)^{\beta/(1-\beta)} \alpha N_0^{\beta}$	$(1 + f)^{\beta/(1-\beta)}N_0$	0	$w0$

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.

$$\frac{dD(E,N)}{dN} > 0. \tag{4.83}$$

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

$$\left(\frac{D(E,N)}{N}\right)^{\gamma} = \left(\frac{E\,N^{\phi}}{N}\right)^{\gamma} = \left(\frac{E^{1/(1-\phi)}}{N}\right)^{\gamma(1-\phi)}. \tag{4.84}$$

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

$$\frac{dN}{dt} = c_1\left[\frac{\pi \alpha N^{\beta}}{N} - \left(\frac{E}{N}\right)^{\gamma} - w_0\right] \qquad (4.29)$$

may be applicable at the national level.

Making the parallel argument for Equation 4.12

$$Q = \alpha N^{\beta} \qquad \beta \in [0,1] \qquad (4.12)$$

---

[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.

Chapter 5: Punishment Policies' Effect on Illicit Drug Users' Purchasing Habits

5.0 Introduction

One often hears proposals to "get tough"[1] on drugs by imposing stiff sanctions for possession of even small amounts that are suitable for personal consumption.[2] Such "zero-tolerance" policies have been criticized because they oblige enforcement agencies to allocate scarce resources to relatively less important offenders and because they violate the principle that "the punishment should fit the crime."

This chapter argues that, in addition, a counter-intuitive phenomenon may occur. That is, consumption may actually increase when punishment increases from a policy in which punishment is proportional to the quantity possessed to a policy in which the maximum possible punishment is imposed irrespective of the quantity possessed. This seems to contradict a fundamental tenet of economics: that consumption should decrease when price increases.

The resolution of this conundrum also comes from elementary economics. Changing a proportional punishment to a maximum punishment policy does in fact increase the average punishment, but the marginal cost decreases because the punishment component of costs is converted to a fixed cost. The theory of the consumer dictates that people equate marginal benefit with marginal cost. So as long as the benefit of using is concave in quantity, as is commonly assumed for consumer goods, decreasing the marginal cost increases consumption.

Section 5.1 refines this argument verbally to develop the reader's intuition for this phenomenon. Section 5.2 introduces a simple mathematical model of the decisions facing a drug user. Section 5.3 solves the model for various punishment policies to determine how they might affect users' purchasing behavior. Section 5.4 compares the solutions and argues that there are benefits to

---

[1] Informal terms will be used when the benefits of being concise seem to outweigh the dangers of not being precise.

[2] For example, J. Robinson (1989) reports that William Bennet, director of the National Office of Drug Control Policy, did so on an NBC News Program in March 1989. More recently, James P. Guy, President of the International Narcotics Enforcement Officers Association, delivered an address before the I.N.E.O.A. Conference in November 1989 in which he asserted that, "We must promote zero tolerance for abusers." Such positions are not confined to the current administration; Lou Canon took this position in an August 28, 1989 piece in the Boston Globe.

making the punishment an increasing function in the quantity possessed at the time of arrest. Section 5.5 generalizes the model by (1) allowing a more general form of the term describing the benefits of using, (2) applying the results to dealers as well as users, and (3) discussing some possible advantages of having different punishment policies for repeat offenders. The final section offers some concluding comments.

5.1 Intuition Behind the Fundamental Result

This chapter develops a model which makes the argument precise, but it is helpful to understand the fundamental dynamics before plunging into the mathematical details. With this in mind, the following paragraphs informally rephrase the argument using three different metaphors.

The first is the adage, "In for a penny, in for a pound." When the punishment for possession does not increase with quantity, this may be good advice. Presumably the user derives some benefit from consuming the drug, and, if drugs are like most consumer products, "more is better." So, if the (punishment) costs are the same whether the user buys a lot or a little, why not buy a lot?

Another perspective comes from comparing the individual's decision about how much to consume to a firm's decision about how much to produce. A firm derives revenue by selling goods produced by employing a variety of factors. Likewise, the individual derives satisfaction by consuming drugs purchased by incurring a variety of costs. Firms maximize profit when marginal revenue equals marginal cost. Likewise, the individual maximizes utility when the marginal benefit of consuming equals the marginal cost of acquiring the drugs.

Changing punishment from an increasing function to a constant independent of the quantity consumed is like converting one of the firm's variable costs to a fixed cost. If the fixed cost is sufficiently high, the firm will cease production, but if it is profitable to produce at all, the profit maximizing quantity is determined by setting marginal revenue equal to marginal cost. Once the decision to produce has been made, fixed costs are sunk costs and do not affect the optimal production level.

Similarly, if punishment is constant for all (positive) quantities, once the individual has decided to use drugs, the level of punishment

has no effect on the optimal amount of consumption.[3] Assuming that "more is better" and there are diminishing returns to consumption, reducing the marginal cost of using will increase the utility maximizing amount of consumption.

The third metaphor also compares the drug user to a firm. For a firm, when the price of an input increases, the optimal mix of factor inputs changes. For example, if wages increase the firm may substitute capital for labor. Production will be lower, but depending on the relative magnitudes of the income and substitution effects,[4] the amount of capital consumed may increase.

The user "produces" a net utility equal to the satisfaction derived from using minus the costs. If the cost of making frequent purchases increases, the user may substitute total quantity for frequency of purchase by increasing the size of each purchase. The user's net utility declines, but the total amount consumed may increase if the substitution effect is greater than the income effect.

It is hoped that these comparisons have given some intuition for the argument explored in this chapter, but they should not be taken too literally. Drug consumption is determined by two interdependent variables: the purchase size and the frequency of purchase, both of which affect the total punishment cost. Hence the response to a change in punishment policy is not as simple as these comparisons suggest, and formalizing the argument with a mathematical model may be useful.

5.2 The Basic Model

5.2.1 Formulation

Mathematical models are inevitably simplifications. Ideally the simplifications make it possible to draw interesting conclusions without so distorting the fundamental nature of the system modelled that the conclusion are erroneous. This section describes the modelling framework used and its underlying assumptions. The assumptions overlook some of the complexity and heterogeneity of

---

[3]Actually this overstates the case. At least in the model developed here, the level of punishment still affects consumption even if it does not depend on the quantity possessed because the frequency of purchase (as well as the purchase size) increases with the consumption rate. The effect can, however, be smaller than it would be if punishment increased with the quantity possessed.

[4]See Varian (1984) for definitions and discussions of the income and substitution effects.

drug users' behavior, but it is hoped that the reader will find that they capture at least most of the first order effects.

A fundamental assumption of the modelling framework used is that drug users maximize their utility. To some it may sound ludicrous to build a model whose foundation rests on the rationality of drug users. After all, their decision to use drugs casts doubt on their foresight, and the use of drugs may cloud whatever judgement the users had originally. Such a view, though commonly held, is not necessarily entirely accurate.

The fact that people have decided to use drugs is not *prima facie* evidence that they are irrational or masochistic. Using drugs involves risks and costs, but so do most activities, and drug use is reported to bring a variety of benefits including, but not necessarily limited to, euphoria, escape, and acceptance in some social groups.[5]

It is also not true that those who have begun using drugs are incapable of acting rationally.[6] First of all, few if any users are continuously "high". Even "hard-core heroin addicts" come down between highs, and many if not most of their purchase decisions are probably made at these times. Secondly, many drug users lead a "normal" life, pursuing a career, raising a family, and maintaining a circle of friends. These include "chippers" who have used drugs for an extended time on a regular basis. Nevertheless, reluctance to describe heavily "addicted" users as people who rationally act in their own self-interest is only reasonable.

Hence, the basic model is restricted to "controlled users." Tautologically this restriction excludes all users whose behavior is inconsistent with one or more of the models' assumptions. For example, it excludes those who do not rationally maximize their individual welfare. It is the author's belief that most users are "controlled users," although they are probably responsible for a smaller fraction of consumption than their numbers might suggest because they consume less on average than compulsive users. Nevertheless, the fact that William Bennett, the director of the Office of National Drug Policy, has frequently spoken out against "controlled users" suggests that they are an important part of "the drug problem."

---

[5]"Usually the rush [of heroin] is described as a violent, orgasmic experience, somewhat like a sexual orgasm, 'only vastly more so'" (Kaplan, 1983a, p.22).

[6]On the contrary, the general impression one receives from reading studies such as Johnson et al. (1985) and Preble and Casey (1969) is that even dependent users behave purposefully and thoughtfully if one remembers that their environment and objectives are not the same as those of the general public.

Besides restricting attention to controlled users, the model focuses on users whose consumption is in "steady state." It does not apply to novice users, people whose consumption varies greatly over relatively short periods, or people who "binge" because modeling the evolution of behavior is complex. This chapter only tries to describe what the ultimate changes might be by examining behavior before and after some exogenous change. In economists' lexicon, it is a comparative statics analysis.

Furthermore, as a considerable simplification, it will be assumed that there is only one kind of drug and that the user buys a particular quantity at regular intervals. Lumping all kinds of drugs together ignores questions about substitution between different kinds of drugs, both licit and illicit. A classic example of this is the possibility that clamping down on marijuana might lead to increased use of potentially more dangerous substances such as PCP.[7] Less direct substitution effects may exist as well. For example, severe punishment for possession of heroin may have enhanced the popularity of cocaine, and widespread use of cocaine may now be fueling demand for heroin for "speedballing."[8] So the model can say nothing about coordinating punishment policies for different drugs.

Likewise using point estimates for the purchase size and frequency is a simplification; clearly consumption is not perfectly regular.

Without these assumptions, however, the analysis below would be considerably more complicated. With them, there are just two decision variables over which the user maximizes utility:

q = the quantity purchased each time the user buys, and
f = the frequency with which the user buys.

Purchasing and using drugs offers a variety of advantages and disadvantages. It will be assumed that these can all be converted into some measure of utility, so the user's decision can be described as an optimization problem:

$$\text{Max } z(f,q) = B(f,q) - C(f,q)$$
$$\text{s.t.} \quad q,f \geq 0$$

where

---

[7]Kleiman (1989, pp.101-102) discusses this possibility.

[8]"Speedballing" is the slang term for using heroin and cocaine together.

B(f,q) = benefit per unit time the user derives from receiving q units of drugs with frequency f, and
C(f,q) = cost per unit time of purchasing q units of drugs at a time with frequency f.

There may be constraints on q and f other than that they be nonnegative. Obviously there is a minimum purchase size suppliers will sell and the frequency of purchase cannot be arbitrarily large or arbitrarily close to zero. Perhaps more significantly, the purchase size q may be limited by the amount of cash the user has.[9] It is assumed that this cash constraint is not binding. For many controlled users this is reasonable. Others may be able to borrow enough to make the desired purchase, and the borrowing costs can be counted as holding costs (discussed below). The problem can also be solved assuming the cash constraint is binding. Then the user will buy as much as he or she can afford as often as possible, and the punishment policy does not affect consumption.

Limited income and wealth also threaten the steady state assumption. One might purchase a fixed quantity q with a fixed frequency f for quite some time, but not be in financial steady state. For example, such an individual might be drawing upon savings.[10] In these situations assuming an infinite horizon steady state may not be objectively realistic, but it may approximate the user's thinking.

There is no "correct" functional form for B(f,q) or C(f,q). After all, even modelling the utility and disutility associated with licit activities that can be studied more easily is imprecise. The best one can do is make some plausible assumptions and hope that the dynamics of the system are relatively robust. The remainder of this subsection lists the assumptions made here.

Assumption A1: $B(f,q) = \alpha \sqrt{fq}$ for $f, q > 0$, where α is a positive constant.

This is the most speculative of the assumptions. For a controlled user who can maintain an inventory (i.e. one who can resist the temptation to binge), B(f,q) is probably a function of the quantity of drugs not f or q individually, i.e. $B(f,q) = B(fq)$. For most consumer goods "more is better" and there are diminishing returns to

---

[9]Some suppliers will sell on credit to familiar customers, but there are limits to the credit they will give.

[10]See, for example, Malcolm, 1989.

consumption, so one would expect $B(fq)$ to be a concave, increasing function.

The square root function is concave and increasing, but it is by no means the only such function. However, other functions, such as $B(fq) = (fq)^{\varepsilon}$ and $B(fq) = \ln(fq)$, lead to less tractable formulations. As will be seen, if $B(f,q) = \alpha \sqrt{fq}$, a number of results can be obtained in closed form. Since there is no obvious reason for preferring any other functional form and it is not feasible to measure the function empirically, the square root function will be used. Section 5.5.1 partially relaxes this assumption, requiring only that $B(fq)$ be increasing and concave, and shows that the overall conclusion holds for this more general case.

C(f,q) includes the following costs: purchase cost, the cost of the negative health effects of using drugs, search cost, the cost of keeping an inventory, and the expected cost of being arrested while buying the drugs. Quantifying these costs requires several assumptions.

Assumption A2: Purchase costs are proportional to the quantity consumed, fq.

Drug markets are certainly large enough that no single user's purchases significantly affect prices. If all users doubled their purchases, prices might increase,[11] but the model focuses on one individual user's decision, so other users' actions are taken as given.

On the other hand, one might object to the assumption because users sometimes obtain quantity discounts.[12] Ignoring these discounts restricts the model to users whose consumption habits clearly place them in the retail market no matter what punishment policy is in effect.

Assumption A3: The costs of the negative health effects are proportional to qf, the quantity consumed.

Since the user is assumed to be in sufficient control of his or her habit to avoid binges, the costs of the negative health effects are probably proportional to the quantity of drugs consumed (qf) not to

---

[11] Chapter 7 discusses the possibility that this would actually make prices decrease because the supply curve is downward sloping.

[12] This is described in Johnson et al. (1985) for heroin users and asserted to be true for cocaine users by Reuter, Crawford, and Cave (1988).

the purchase quantity (q) alone. To avoid cluttering notation, these costs are incorporated into the purchase cost term.

Assumption A4: Search costs are proportional to f.

Search costs for the first few purchases are likely to be much greater than for subsequent purchases, but after that initial transient, search costs, or at least expected search costs, are probably proportional to f for wide ranges in q. The model is of steady state behavior so it ignores any such transient effects and assumes search costs are proportional to f.

Assumption A5: Inventory costs are proportional to q.

Quantities held for personal use take up so little physical space one might think inventory costs would be negligible. In fact, inventory costs are likely to be small relative to other costs, but if they were truly zero, then users would buy all they ever plan to use in one purchase. Clearly this is not done in practice, so there must be disincentives to holding large inventories.

The first of these are simple storage costs. Storage costs per unit weight of drugs are high relative to those of most other consumer products because most users make at least some effort to hide them. Second, there is the risk of having the drugs stolen or inadvertently damaged. Third, there is the opportunity cost of holding inventory; money tied up in an inventory of drugs is money that cannot be used elsewhere. Fourth, at some time most users stop using drugs, and the salvage value of an unused inventory of drugs is next to zero. And finally, at least for marijuana, the quality of the drugs may deteriorate over time.[13]

These costs are proportional to the quantity held. Suppose the user consumes at a constant rate and replenishes whenever the stock drops below some threshold. Then inventory as a function of time has a sawtooth graph and the average inventory cost per unit time is a constant, which can be ignored, plus a term proportional to q.

Assumption A6: The probability of being arrested is proportional to f, the frequency with which purchases are made.

---

[13]According to Kleiman (1989, p.40), "Under uncontrolled conditions, THC content decreases with a half-life measured in months."

This assumption reflects the observation that most users who are arrested are apprehended while buying or when they have the drugs in their possession soon after buying. It is less common to be arrested while using or for possession at some other time.

If every user purchased more often, then the risk of arrest per purchase might decrease because the ratio of police power to the number of transactions would fall. But this model focuses on one individual's decision, and one individual's actions will not appreciably affect this ratio.

Factors other than the amount of police power per transaction may have an effect. Frequenting drug markets might arouse suspicion, so the probability of arrest per purchase might increase with f. On the other hand, frequent purchasers may be more adept at avoiding arrest. Or, it is possible that frequent purchasers may become careless. Also, infrequent purchasers may be able to take better advantage of variations in enforcement pressure. On the other hand, frequent purchasers may know more about these variations. Since the relative strength of these conflicting factors is not obvious, in this paper it will be assumed that the probability of arrest per purchase is independent of the frequency with which purchases are made.

Assumption A7: The expected penalty the user suffers if arrested while buying depends on the quantity purchased.

The expected penalty to the user if he or she is arrested will be represented by the function $c_a(q)$. To a large extent this function, referred to in the sequel as the punishment policy, can be set by policy makers.

It is reasonable to assume punishment depends on the quantity possessed at arrest (especially since the constant function is not excluded), but it may depend on other factors as well. In particular, punishment frequently depends on the offender's prior criminal record, including their record for drug offenses. This suggests that during steady state periods the criminal record must be constant. For example, suppose the individual is arrested for possession of narcotics but released on parole. The expected punishment for that individual upon a subsequent arrest would probably increase, so a new equilibrium would be obtained.

It is possible that a dynamic purchasing strategy would be superior. For example, it may be optimal for the purchase amount to increase slowly as the amount of time since the most recent arrest increases, but such possibilities are ignored. The omission of users'

criminal records is more problematic when one discusses the model's implications for optimal punishment policies.

Section 5.6.2 will discuss one aspect of making punishment depend on arrest histories.

Having made these assumptions the cost function may be modelled as

$$C(f,q) = h\,q + c_s\,f + p\,c_a(q)\,f + c_p\,qf \qquad (5.1)$$

where

h = inventory cost coefficient,
c_s = search time cost per purchase,
$c_a(q)$ = expected cost of being arrested while buying,
p = probability of being arrested while making a purchase, and
c_p = purchase price.

5.2.2 Solution to the Basic Model With Linear Punishment

Given the assumptions above, if $c_a(q)$ is assumed to be linear ($c_a(q) = a + bq$ for $q > 0$) the cost term is

$$C(f,q) = h\,q + (c_s + pa)\,f + (c_p + pb)\,qf \qquad (5.1A)$$

so the user's problem is to

$$\text{Max } z(f,q) = \alpha\sqrt{fq} - h\,q - (c_s + p\,a)\,f - (c_p + p\,b)\,fq \qquad (P1A)$$
$$\text{s.t.} \quad q,f \geq 0.$$

This is a relatively straightforward nonlinear optimization problem in two variables with nonnegativity constraints. The same problem with different coefficients will arise several times, so it convenient to solve a slightly more general form.

$$\text{Max } z(f,q) = \alpha\sqrt{fq} - \beta\,q - \gamma\,f - \delta\,fq \qquad (P\ 1)$$
$$\text{s.t.} \quad q,f \geq 0.$$

The first order conditions are

$$\nabla z(f,q) = \begin{bmatrix} \frac{\partial z(f,q)}{\partial f} \\ \frac{\partial z(f,q)}{\partial q} \end{bmatrix} = \begin{bmatrix} \frac{\alpha}{2}\sqrt{\frac{q}{f}} - \gamma - \delta q \\ \frac{\alpha}{2}\sqrt{\frac{f}{q}} - \beta - \delta f \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \qquad (5.2)$$

The Hessian

$$\nabla^2 z(f,q) = \begin{bmatrix} -\frac{\alpha}{4} q^{\frac{1}{2}} f^{\frac{-3}{2}} & \frac{\alpha}{4} q^{\frac{-1}{2}} f^{\frac{-1}{2}} - \delta \\ \frac{\alpha}{4} q^{\frac{-1}{2}} f^{\frac{-1}{2}} - \delta & -\frac{\alpha}{4} f^{\frac{1}{2}} q^{\frac{-3}{2}} \end{bmatrix} \tag{5.3}$$

is negative definite because its first principal minor is less than zero,

$$P_1 = -\frac{\alpha}{4} q^{\frac{-1}{2}} f^{\frac{-3}{2}} < 0 \tag{5.4}$$

and its second principal minor

$$P_2 = \left(\frac{\alpha}{4\sqrt{qf}}\right)^2 - \left(\frac{\alpha}{4\sqrt{qf}} - \delta\right)^2 > 0 \tag{5.5}$$

is positive. Hence, z(f,q) is concave, and solving the first order conditions gives the unique global maximum. Let * denote an optimal value. If

$$\alpha \leq 2\sqrt{\beta\gamma} \tag{5.6}$$

the solutions to the first order conditions are negative so $z^* = f^* = q^* = 0$. But if $\alpha > 2\sqrt{\beta\gamma}$, then

$$q^* = \frac{\alpha}{2\delta}\sqrt{\frac{\gamma}{\beta}} - \frac{\gamma}{\delta}, \tag{5.7}$$

$$f^* = \frac{\beta}{\gamma} q^* = \frac{\alpha}{2\delta}\sqrt{\frac{\beta}{\gamma}} - \frac{\beta}{\delta}, \tag{5.8}$$

$$q^* f^* = \left(\frac{\frac{\alpha}{2} - \sqrt{\beta\gamma}}{\delta}\right)^2, \text{ and} \tag{5.9}$$

$$z(f^*, q^*) = \delta q^* f^* = \frac{\left(\frac{\alpha}{2} - \sqrt{\beta\gamma}\right)^2}{\delta}. \tag{5.10}$$

Substituting h for β, $(c_s + pa)$ for γ, and $(c_p + pb)$ for δ, shows that for Problem P1A if

$$\alpha \le 2\sqrt{h(c_s + p\,a)} \tag{5.6A}$$

the solution is $z^* = f^* = q^* = 0$, but if $\alpha > 2\sqrt{h(c_s + p\,a)}$ the optimal purchase quantity and frequency are

$$q^* = \frac{\alpha}{2(c_p + p\,b)}\sqrt{\frac{c_s + p\,a}{h}} - \frac{c_s + p\,a}{c_p + p\,b} \quad \text{and} \tag{5.7A}$$

$$f^* = \frac{h}{c_s + p\,a}q^* = \frac{\alpha}{2(c_p + p\,b)}\sqrt{\frac{h}{c_s + p\,a}} - \frac{h}{c_p + p\,b} \tag{5.8A}$$

and the optimal rate of consumption and objective function value are

$$q^* f^* = \left(\frac{\frac{\alpha}{2} - \sqrt{h(c_s + p\,a)}}{c_p + p\,b}\right)^2 \quad \text{and} \tag{5.9A}$$

$$z(f^*, q^*) = (c_p + p\,b)q^* f^* = \frac{\left(\frac{\alpha}{2} - \sqrt{h(c_s + p\,a)}\right)^2}{c_p + p\,b}\,. \tag{5.10A}$$

These expressions look more complicated than they are. Consider, for example, the condition that if $\alpha \le 2\sqrt{h(c_s + p\,a)}$ the user will not purchase or use drugs. Since α is proportional to the benefit derived from using, it makes sense that if α is small the costs of using might outweigh the benefits. Likewise, the holding cost parameter (h), search cost parameter (c_s), probability of arrest (p), and intercept of the punishment function (a) are all measures of cost. The higher they are, the more likely it is that α is not large enough for it to be worthwhile to consume. (The term $(c_p + pb)$ does not appear in this condition, because it is the coefficient of a higher order term that is negligible for small q and f.)

Consider now the case when it is optimal to consume. The optimal purchase quantity q*, frequency of purchase f*, consumption rate q*f*, and utility z(f*, q*) are all increasing in α and are generally

decreasing in all the cost parameters, as would be expected. The exceptions are that f* can be increasing in h, the coefficient of q in the objective function, and q* can be increasing in (c_s + pa), the coefficient of f in the objective function. This is not surprising since q and f are substitutes. They may be increasing or decreasing functions of the other's cost coefficient depending on whether the income or substitution effect dominates.[14]

Note f* is proportional to q*, and the proportionality constant is the ratio of their cost coefficients. This is because everywhere else in the objective function q and f appear together as qf. Hence, to first order, reducing q by 1% can be compensated for by increasing f by q*/f* per cent, so the price of q (h) must be equal to q*/f* times the price of f (c_s + pa), i.e. $f^* = \frac{h}{c_s + p\,a} q^*$.

Not surprisingly people for whom search costs are low but holding costs are high will purchase smaller quantities more frequently. Compare, for instance, a stereotypical addict with an affluent user. Search costs for addicts are probably lower because they are more likely to live near a drug market and are less likely to be mugged when they go into that neighborhood. On the other hand, addicts' holding costs are probably higher for at least three reasons. First, they are more likely to binge. Second, addicts are more likely to be cash constrained. Third, and perhaps most importantly, the addict's personal inventory is more likely to be stolen. Addicts living on the street are very vulnerable to robbery. If they carry anything of value they are likely to lose it and perhaps life or limb as well. In contrast, if an affluent user keeps $500 worth of drugs at home it will have almost no impact on the chance he or she will be burglarized. Most burglars would not know the drugs were there and there are already other valuable items in the house. All of this suggests that affluent users are more likely than street addicts to buy larger quantities less frequently.

Equations 5.6A-5.10A have several policy implications. For instance, since the punishment policy parameters a and b always appear as pa and pb, if the probability of arrest (p) is small, increasing the expected punishment (increasing a or b or both) will have little effect.

---

[14]The substitution effect dominates, and thus q* and f* are increasing in the other's cost parameter, if the user is a heavy user, more specifically, if $\alpha > 4\sqrt{h(c_s + p\,a)}$.

In that case the policy maker may be forced to work with the search time cost c_s and purchase price c_p. (parameters α and h probably cannot be influenced directly.) If the user is a "committed user" who derives great satisfaction from consuming drugs, i.e. $\alpha >> 2\sqrt{h(c_s + p\,a)}$, then increasing search time costs will not have much effect compared with increasing prices. Increasing prices will also reduce the consumption of users who are "on the fence" ($\alpha \approx 2\sqrt{h(c_s + p\,a)}$), but it will never push them out of the market completely. Increasing search time costs enough to make $\alpha \leq 2\sqrt{h(c_s + p\,a)}$ may, however, do exactly that.

This directly supports Moore's classic argument about the benefits of price discrimination in the retail heroin market.[15] Specifically, the best way to discourage novice users is to raise search time costs, not to raise prices. On the other hand, it suggests that increasing search time costs will have less effect on heavy users because they can adapt by purchasing larger quantities less often. Some analysts suspect that this occurs, so many final sales are for more than the retail quantities.[16]

5.2.3 Setting the Punishment Policy

Although enforcement policy affects price and search time costs (and perhaps α and h less directly), $c_a(q)$ is all that can be controlled directly, so it is natural to ask what it should be. The answer depends, of course, on one's objectives. This paper will assume the objective is to minimize the rate of consumption (q^*f^*) which, if price is constant, is also proportional to the drug dealers' revenues.

From the solution above, if $\alpha \leq 2\sqrt{h(c_s + p\,a)}$, $q^* = f^* = 0$. Rearranging this expression shows that if $a \geq \frac{1}{p}\left(\frac{\alpha^2}{4\,h} - c_s\right)$, $q^* = f^* = 0$ for any b. Similarly, if $\alpha > 2\sqrt{h(c_s + p\,a)}$, $q^* f^* = \left(\frac{\frac{\alpha}{2} - \sqrt{h(c_s + p\,a)}}{c_p + p\,b}\right)^2$, so for any given a, q^* and f^* can be made arbitrarily small by choosing b large enough. In other words, according to this simple model, if the expected punishment (not just the threatened punishment) were made sufficiently severe, drug consumption could be eliminated.

---

[15]Moore, 1977.

[16]Reuter and Kleiman (1986, p.295) and Reuter, Crawford, and Cave (1988, p.23).

There are limits, however, to the severity of punishment that society will tolerate. It is unlikely, for example, that the death penalty (imposed by slow immersion in boiling oil) will ever be instituted for the possession of trace amounts of marijuana. Let c_a be the maximum punishment for a drug user that society will tolerate.[17] Then the constraint $c_a(q) \leq c_a$ should be added to the formulation. Society's sense of justice would probably also demand that $c_a(q)$ be a nondecreasing function of q and $c_a(0) = 0$.

5.3 Evaluating Various Punishment Policies

5.3.1 Policy of Maximum Punishment

One might think that imposing the harshest possible penalties ($c_a(q) = c_a$ for all $q > 0$) would minimize consumption. With such a maximum punishment policy the cost term is

$$C(f,q) = h\,q + c_s f + p\,c_a f + c_p qf \qquad (5.1B)$$

and the user's optimization problem is

$$\text{Max } z(f,q) = \alpha\sqrt{fq} - h\,q - (c_s + p\,c_a) f - c_p fq \qquad (P1B)$$
$$\text{s.t.} \quad q,f \geq 0.$$

This is the special case of Problem P1 with $a = c_a$ and $b = 0$. The solution can be found by substituting those values into the expressions in Section 5.2.2. Capital letters will be used to denote the optimal solution under a maximum punishment policy.

If

$$\alpha \leq 2\sqrt{h(c_s + p\,c_a)} \qquad (5.6B)$$

then $F = Q = 0$, but if $\alpha > 2\sqrt{h(c_s + p\,c_a)}$, the optimal purchase quantity and frequency are

$$Q = \frac{\alpha}{2c_p}\sqrt{\frac{c_s + p\,c_a}{h}} - \frac{c_s + p\,c_a}{c_p} \text{ and} \qquad (5.7B)$$

---

[17] c_a can be drug specific if the model is interpreted as applying to just one drug and substitution is ignored.

$$F = \frac{h}{c_s + p\,c_a} Q = \frac{\alpha}{2c_p}\sqrt{\frac{h}{c_s + p\,c_a}} - \frac{h}{c_p}, \qquad (5.8B)$$

and the optimal rate of consumption and objective function value are

$$Q\,F = \left(\frac{\frac{\alpha}{2} - \sqrt{h(c_s + p\,c_a)}}{c_p}\right)^2 \quad \text{and} \qquad (5.9B)$$

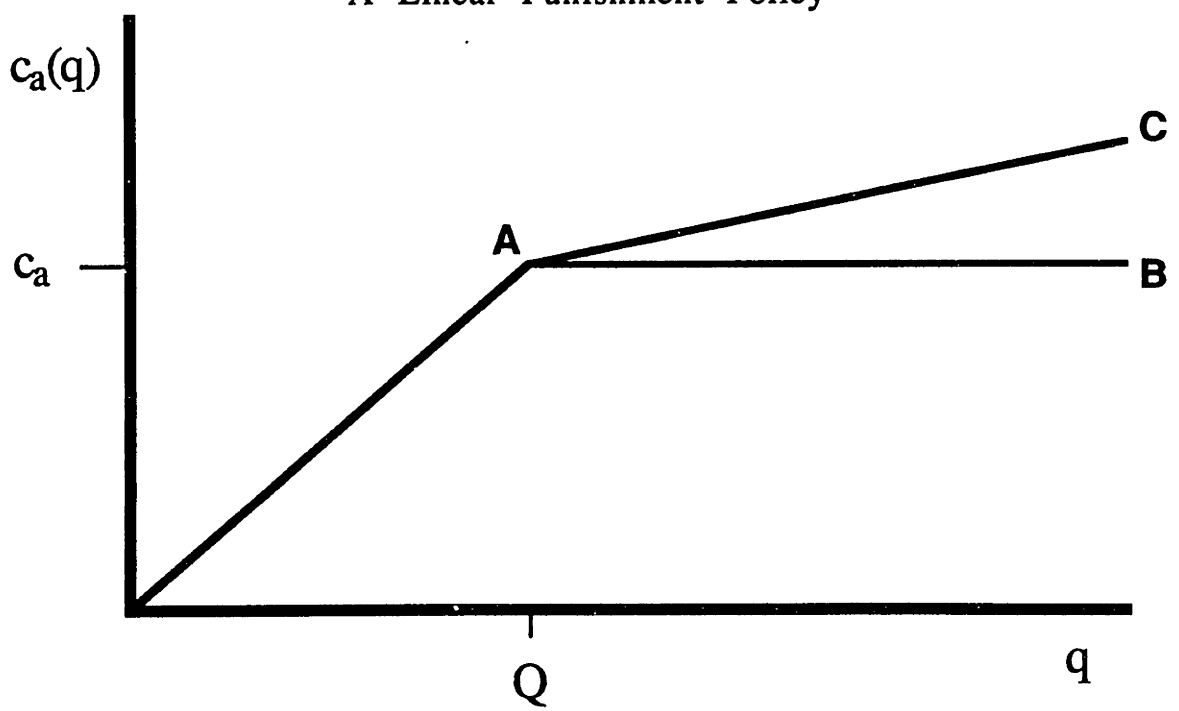
$$z(F,Q;\, c_a(q) = c_a) = c_p\,Q\,F = \frac{\left(\frac{\alpha}{2} - \sqrt{h(c_s + p\,c_a)}\right)^2}{c_p}. \qquad (5.10B)$$

5.3.2 A Policy That Leads to Less Consumption

Surprisingly, this maximum punishment policy does not minimize consumption. This section shows the rate of consumption will be lower if the expected cost of being arrested increases linearly with q from 0 to c_a for $0 \le q \le Q$ and is no less than c_a for $q \ge Q$. (See Figure 5.1.) That is, if

$$c_a(q) = \frac{c_a}{Q} q \qquad \text{for } 0 \le q \le Q, \text{ and} \qquad (5.11)$$
$$c_a(q) \ge c_a \qquad \text{for } q \ge Q.$$

Figure 5.1:
A Linear Punishment Policy



The exact form of $c_a(q)$ for $q > Q$ does not matter because the user will never possess a quantity $q > Q$. To see this, look again at Figure 5.1. The previous section showed the user prefers point A to any point on the line (A,B). But for a fixed q increasing the punishment cannot improve the user's utility, so the user will prefer any point on (A,B) to the corresponding point on (A,C) directly above. Since A is feasible under the new policy, this means the user will never pick any point on (A,C), i.e. $q^* \leq Q$.

More formally, let $z(f, q; \frac{c_a}{Q}q)$ denote the user's utility function with this punishment policy. For all f and all $q \geq Q$, $z(f, q; \frac{c_a}{Q}q) \leq z(f, q; c_a(q) = c_a) \leq z(F, Q; c_a(q) = c_a) = z(F, Q; \frac{c_a}{Q}q)$. This implies there is an optimal solution with $q^* \leq Q$.

Thus the exact form of $c_a(q)$ for $q > Q$ is irrelevant. In particular, the same solution will be obtained if $c_a(q) = \frac{c_a}{Q} q$ for all $q \geq 0$. This is the special case of Problem P1 with $a = 0$ and $b = \frac{c_a}{Q}$ so the solution can be obtained by substituting these values into the expressions in Section 5.2.2. Let ~ denote quantities that are optimal with this punishment policy. If

$$\alpha \leq 2\sqrt{h\, c_s} \tag{5.6C}$$

then $\tilde{q} = \tilde{f} = 0$, but if $\alpha > 2\sqrt{h\, c_s}$ the solution is

$$\tilde{q} = \frac{\alpha}{2\left(c_p + \frac{p\, c_a}{Q}\right)}\sqrt{\frac{c_s}{h}} - \frac{c_s}{\left(c_p + \frac{p\, c_a}{Q}\right)}, \tag{5.7C}$$

$$\tilde{f} = \frac{\alpha}{2\left(c_p + \frac{p\, c_a}{Q}\right)}\sqrt{\frac{h}{c_s}} - \frac{h}{\left(c_p + \frac{p\, c_a}{Q}\right)}, \tag{5.8C}$$

$$\tilde{q}\,\tilde{f} = \left(\frac{\frac{\alpha}{2} - \sqrt{h\,c_s}}{\left(c_p + \frac{p\,c_a}{Q} \right)} \right)^2, \quad \text{and} \tag{5.9C}$$

$$z(\tilde{f}, \tilde{q}; \tfrac{c_a}{Q} q) = \left(c_p + \frac{p\,c_a}{Q} \right) \tilde{q}\,\tilde{f} = \frac{\left(\frac{\alpha}{2} - \sqrt{h\,c_s} \right)^2}{c_p + \frac{p\,c_a}{Q}} . \tag{5.10C}$$

Individuals with $2\sqrt{h\,c_s} < \alpha < 2\sqrt{h\,(c_s + p\,c_a)}$ consume a positive quantity under this punishment policy but not under a policy of maximum punishment. This suggests that switching from a maximum punishment policy to this policy might increase the number of users. This is to be expected because curious people are more likely to experiment if they know they will not be punished as severely as heavy users.

If $\alpha > 2\sqrt{h\,(c_s + p\,c_a)}$, however, switching from a maximum punishment policy has a very different result. In that case the average purchase size is smaller with the linear punishment policy since

$$Q - \tilde{q} = \frac{\alpha\,Q}{2\sqrt{h}\,(c_p\,Q + p\,c_a)} \left(\sqrt{c_s + p\,c_a} - \sqrt{c_s} \right) > 0. \tag{5.12}$$

Also, it must be the case that the user is better off because he or she could have q = Q, f = F, and risk the same punishment as with the maximum punishment policy. Since the previous solution is feasible and the user maximizes utility, the new optimal solution's utility must be at least as great. In symbols,

$$z(\tilde{f}, \tilde{q}; \tfrac{c_a}{Q} q) \geq z(F, Q; \tfrac{c_a}{Q} q) = z(F, Q; c_a(q) = c_a). \tag{5.13}$$

Actually the inequality is strict because

$$z(\tilde{f}, \tilde{q}; \tfrac{c_a}{Q} q) - z(F, Q; c_a(q) = c_a)$$

$$= \frac{\alpha\left(\frac{\alpha}{2} - \sqrt{h\,(c_s + p\,c_a)}\right)}{c_p\,(c_pQ + p\,c_a)} \left(\sqrt{\frac{c_s + p\,c_a}{2}} - \sqrt{\frac{c_s}{2}}\right)^2 > 0. \qquad (5.14)$$

The user will buy more frequently since

$$\tilde{f} - F = \frac{\frac{\alpha}{2} - \sqrt{h\,(c_s + p\,c_a)}}{2\,c_p\,(c_pQ + p\,c_a)}\left[\alpha\left(\sqrt{\frac{c_s + p\,c_a}{c_s}} - 1\right) + \sqrt{\frac{h}{c_s + p\,c_a}}\,2\,p\,c_a\right] > 0. \qquad (5.15)$$

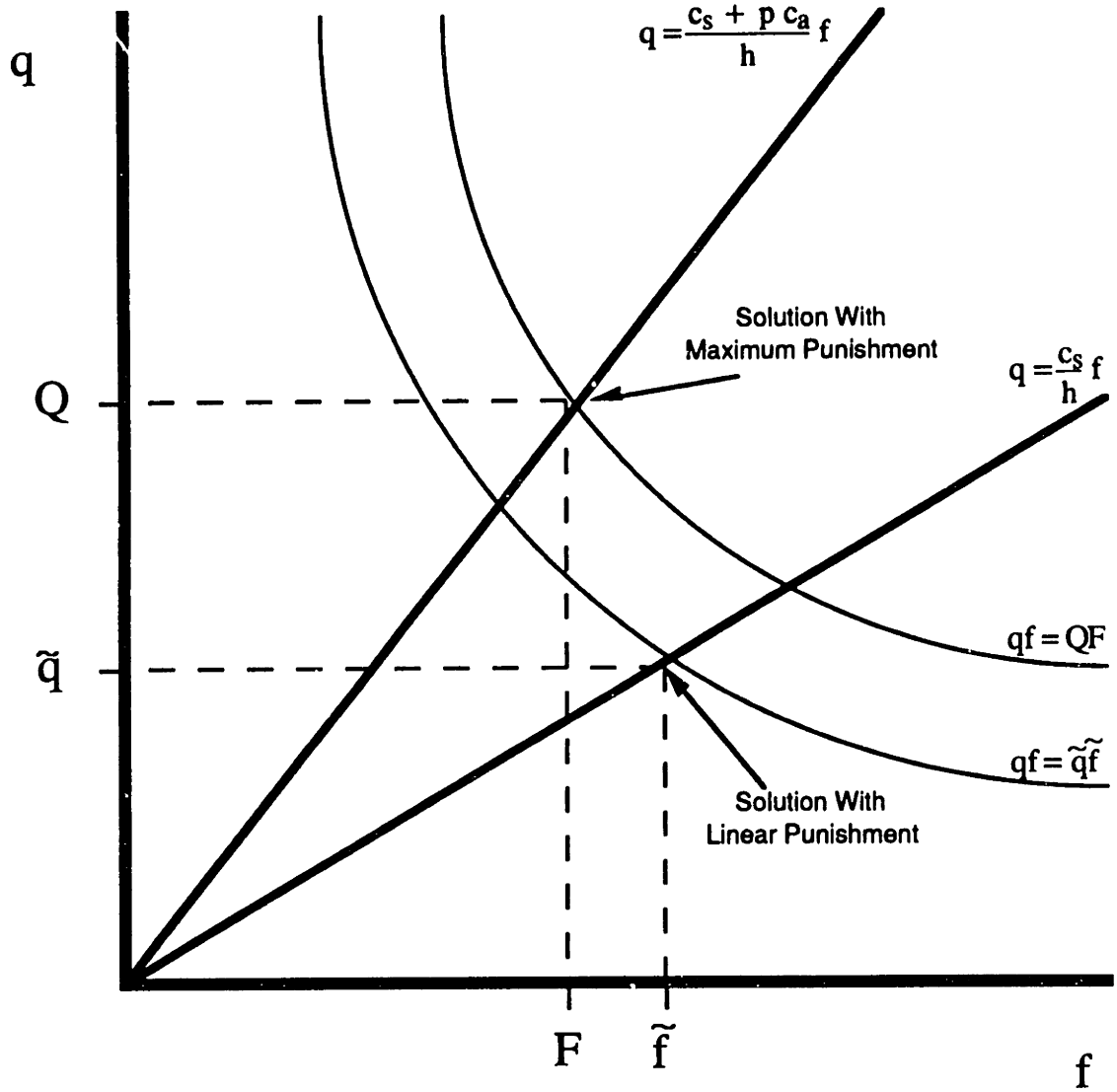
Surprisingly, the size of the average purchase decreases enough to more than counteract the increased frequency with which purchases are made, so the overall rate of consumption decreases by

$$QF - \tilde{q}\tilde{f} = \left(\frac{\frac{\alpha}{2} - \sqrt{h(c_s + p\,c_a)}}{c_p}\right)^2 - \left(\frac{\frac{\alpha}{2} - \sqrt{h\,c_s}}{\left(c_p + \frac{p\,c_a}{Q}\right)}\right)^2 \qquad (5.16)$$

$$= \frac{(c_pQ)^2}{(2\,c_p\,(c_pQ + p\,c_a))^2}\left[\left(\alpha - 2\sqrt{\frac{c_s}{c_s + p\,c_a}}\sqrt{h\,c_s}\right)^2 - \left(\alpha - 2\sqrt{h\,c_s}\right)^2\right] > 0.$$

Figure 5.2 shows these changes in the q-f plane.

Figure 5.2:
Changing from Maximum Punishment to a Linear Policy



5.3.3 Consumption-Minimizing Policy

The previous section showed that a maximum punishment policy need not minimize consumption. A natural question to ask is, therefore, what punishment policy will minimize consumption?

This question can also be posed as an optimization problem, but now the argument of the minimization is the function $c_a(q)$. That is, the decision maker specifies an entire function, not just a finite number of variables.

Not every function is feasible because the punishment policy must be consistent with society's sense of justice. Specifically, it seems reasonable to restrict attention to policies that do not punish people who do not have any drugs ($c_a(0) = 0$), that never threaten more than the maximum permissible punishment ($c_a(q) \le c_a$), and that punish people possessing large quantities at least as severely as those carrying smaller amounts ($c_a(q)$ is nondecreasing).

The policy maker would like to find the punishment policy $c_a(q)$ satisfying these constraints that minimizes the rate of consumption, qf. But q and f both depend on the punishment policy. They are determined when the user maximizes his or her utility, and as the preceding discussion showed, the user's optimal purchasing pattern is affected by the punishment policy.

Mathematically this can be represented as a nested optimization problem:

$$\underset{c_a(q) \in F}{\text{Min}} \left\{ f^* q^* \mid (f^*, q^*) = \arg \underset{f,q \ge 0}{\text{Max}} \left\{ z(f,q) = \alpha\sqrt{fq} - hq - (c_s + p\, c_a(q))\, f - c_p\, f\, q \right\} \right\} \quad \text{(P 2)}$$

where $F = \{f: \Re \to \Re \mid f(0) = 0,\ f(q) \le c_a\ \forall q,$ and $f()$ is nondecreasing$\}$ is the set of feasible punishment policies.

The internal maximization represents the user's problem of maximizing utility subject to the punishment policy $c_a(q)$. The outer minimization represents the policy maker's problem of choosing a punishment policy that minimizes the rate of consumption, qf.

The derivation of the solution is more technical than the rest of the chapter, so readers may want to skip the introductory lemmas and proceed directly to the theorem at the end of this subsection.

Setting the partial derivative of z(f,q) with respect to f equal to zero

$$\frac{\partial z(f,q)}{\partial f} = \frac{\alpha}{2}\sqrt{\frac{q}{f}} - \left(c_s + p c_a(q) + c_p q\right) = 0 \qquad (5.17)$$

shows that

$$f^* = \frac{\frac{1}{4}\alpha^2 q^*}{\left(c_s + pc_a^*(q^*) + c_p q^*\right)^2}. \qquad (5.18)$$

Substituting this into $z(f,q) = \alpha\sqrt{fq} - hq - (c_s + p\,c_a(q))f - c_p f q$ implies that

$$z(f^*, q^*) = \frac{\alpha^2 q^*}{4\left(c_S + p\,c_a^*(q^*) + c_p q^*\right)} - h\,q^*. \qquad (5.19)$$

Lemma 1: $q^* \le Q = \frac{\alpha}{2c_p}\sqrt{\frac{c_s + p\,c_a}{h}} - \frac{c_s + p\,c_a}{c_p}$, the optimal quantity when $c_a(q) = c_a$ for all $q > 0$.

Proof: Suppose to the contrary that $q^* > Q$. Since $c_a(q) = c_a$ is in the set of feasible punishment policies

$$q^* f^* = \frac{\alpha^2 q^{*2}}{4\left(c_S + p\,c_a^*(q^*) + c_p q^*\right)^2} \le \frac{\alpha^2 Q^2}{4\left(c_S + p\,c_a + c_p Q\right)^2} = QF.$$

Since $c_a^*(q)$ must also be feasible, $c_a^*(q^*) \le c_a$. So

$$\frac{\alpha^2 Q^2}{4\left(c_S + p\,c_a + c_p Q\right)^2} \le \frac{\alpha^2 Q^2}{4\left(c_S + p\,c_a^*(q^*) + c_p Q\right)^2},$$

which implies

$$\frac{q^*}{c_S + p\,c_a^*(q^*) + c_p \hat{q}} \le \frac{Q}{c_S + p\,c_a^*(q^*) + c_p Q}.$$

Thus since $c_s + pc_a^*(q^*) > 0$ and $c_p > 0$, $q^* \le Q$. Contradiction. QED.

Lemma 2: There is an optimal punishment policy with $c_a^*(Q) = c_a$.

Proof: By Lemma 1, $q^* \le Q$. If $q^* = Q$ then $c_a^*(Q) = c_a$ minimizes f^* and thus q^*f^*.

Suppose $q^* < Q$ and $c_a^{**}(q)$ is an optimal punishment policy with $c_a^{**}(Q) < c_a$. Then the punishment policy

$$c_a^*(q) = \begin{cases} c_a^{**}(q) & q < Q \\ c_a & q \geq Q \end{cases}$$

yields the same solution as $c_a^{**}(q)$ does because $z(f, q; c_a^*(q)) = z(f, q; c_a^{**}(q))$ for all $q < Q$ and $z(f, q; c_a^*(q)) \leq z(f, q; c_a^{**}(q)) \leq z(f^*, q^*; c_a^{**}(q^*)) = z(f^*, q^*; c_a^*(q))$ for all $q \geq Q$. QED.

Lemma 3: There is an optimal punishment policy $c_a^*(q)$ and quantity q^* such that $c_a^*(q^*) \leq U(q^*)$ where

$$pU(q) = \frac{\frac{1}{4}\alpha^2 q c_p}{\left(\frac{\alpha}{2} - \sqrt{h(c_s + pc_a)}\right)^2 + h q c_p} - - c_s - c_p q .$$

Proof: By Lemma 2 there exists an optimal punishment policy with $c_a^*(Q) = c_a$. Since (F,Q) is feasible for the user, $z(f^*, q^*; c_a^*(q)) \geq z(F, Q; c_a^*(q)) = z(F, Q; c_a(q) = c_a)$, i.e.

$$z\left(f^*, q^*; c_a^*(q)\right) = \frac{\alpha^2 q^*}{4\left(c_S + p\, c_a^*(q^*) + c_p q^*\right)} - h q^* \geq \frac{\left(\frac{\alpha}{2} - \sqrt{h(c_S + p c_a)}\right)^2}{c_p} .$$

Solving for $pc_a^*(q^*)$ gives the desired inequality. QED.

Theorem 1: Any punishment policy that satisfies the following conditions will minimize consumption.

1) $c_a^*(q) = 0$ for $0 \leq q \leq \bar{q}$,
2) $c_a^*(q) \geq U(q)$ for $\bar{q} \leq q \leq Q$, and
3) $c_a^*(q) = c_a$ for $q \geq Q$

where

$$U(q) = \frac{1}{p}\left[\frac{\frac{1}{4}\alpha^2 q c_p}{\left(\frac{\alpha}{2} - \sqrt{h(c_s + pc_a)}\right)^2 + h q c_p} - c_s - c_p q\right] \text{ and } \qquad (5.20)$$

$Q = \frac{\alpha}{2c_p}\sqrt{\frac{c_s + p\,c_a}{h}} - \frac{c_s + p\,c_a}{c_p}$ (c.f Equation 5.7B) is the optimal purchase quantity when $c_a(q) = c_a$. With any consumption-minimizing policy:

$$\bar{q} = \frac{1}{c_p}\left[\frac{\alpha}{2}\sqrt{\frac{c_s + p\,c_a}{h}} - c_s + \frac{p\,c_a}{2} - \frac{\sqrt{p\,c_a}}{2}\sqrt{p\,c_a + \frac{\alpha}{h}\left(\alpha - 2\sqrt{h\,(c_s + p\,c_a)}\right)}\right], \quad (5.7D)$$

$$\bar{f} = \frac{\alpha^2\bar{q}}{4\left(c_s + c_p\bar{q}\right)^2}, \quad (5.8D)$$

$$\bar{q}\,\bar{f} = \frac{\alpha^2\bar{q}^2}{4\left(c_s + c_p\bar{q}\right)^2} = \left(\frac{\alpha\bar{q}}{2\left(c_s + c_p\bar{q}\right)}\right)^2, \text{ and} \quad (5.9D)$$

$$z\left(\bar{f}, \bar{q}; c_a^*(q)\right) = \frac{\left(\frac{\alpha}{2} - \sqrt{h\,(c_s + p\,c_a)}\right)^2}{c_p}. \quad (5.10D)$$

Proof: Note that U(q) is an indifference curve for the user since $z(f, q; c_a(q) = U(q)) = z(F, Q; c_a(q) = c_a)$. It has the following properties: $U(Q) = c_a$, $\frac{d\,(U(q=0))}{dq} > 0$, $\frac{d\,(U(q=Q))}{dq} = 0$, and $\frac{d^2(U(q))}{dq^2} < 0$ so U(q) is increasing for $0 \le q \le Q$. Also U(q) = 0 for

$$q = Q + \frac{p\,c_a}{2\,c_p} \pm \frac{\sqrt{p\,c_a}}{2\,c_p}\sqrt{p\,c_a + \frac{\alpha}{h}\left(\alpha - 2\sqrt{h\,(c_s + p\,c_a)}\right)}.$$

Let $\bar{q}$ denote the smaller of the two zeros. Then

$$c_p\,\bar{q} = \frac{\alpha}{2}\sqrt{\frac{c_s + p\,c_a}{h}} - c_s + \frac{p\,c_a}{2} - \frac{\sqrt{p\,c_a}}{2}\sqrt{p\,c_a + \frac{\alpha}{h}\left(\alpha - 2\sqrt{h\,(c_s + p\,c_a)}\right)}$$

and $0 < \bar{q} < Q$.

Now Lemma 3 implies $q^* \geq \bar{q}$; otherwise, $c_a^*(q^*)$ would be negative. Think of the problem in the quantity-punishment plane. The solution $(q^*, c_a^*(q^*))$ lies in the region bounded by $\bar{q} \leq q^* \leq Q$ and $0 \leq c_a^*(q^*) \leq U(q^*)$. Along the curve $U(q^*)$ for $\bar{q} \leq q^* \leq Q$,

$$q^* f^* = \frac{\alpha^2 q^{*2}}{4\left(c_S + p\,U(q^*) + c_p q^*\right)^2} = \frac{\frac{1}{4}\alpha^2 q^{*2}}{\left(\frac{\frac{1}{4}\alpha^2 c_p q^*}{\left(\frac{\alpha}{2} - \sqrt{h\left(c_S + p\,c_a\right)}\right)^2 + h c_p q^*}\right)^2} = \frac{\left(\left(\frac{\alpha}{2} - \sqrt{h\left(c_S + p\,c_a\right)}\right)^2 + h c_p q^*\right)^2}{\frac{1}{4}\alpha^2 c_p^2}.$$

This is strictly increasing in q^*, so it is minimized along this curve by taking $q^* = \bar{q}$.

Since

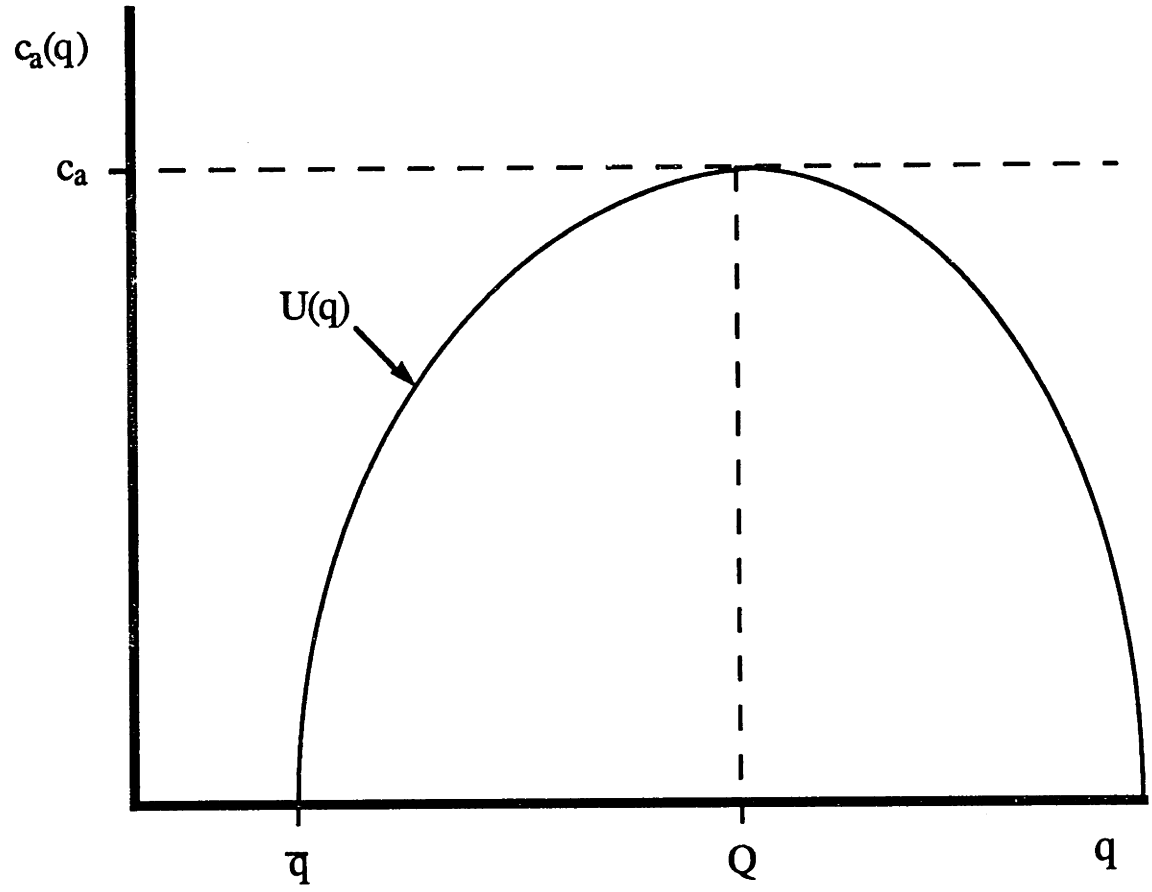
$$q^* f^* = \frac{\frac{1}{4}\alpha^2 q^*}{\left(c_S + p c_a^*(q^*) + c_p q^*\right)^2}$$

no point $(q^*, c_a^*(q^*))$ in the region can be better than the corresponding point $(q^*, U(q^*))$ on the curve $U(q^*)$. Hence, a policy that forces $q^* = \bar{q}$ is optimal, and any $c_a^*(q)$ such that $c_a^*(q) = 0$ for $0 \leq q \leq \bar{q}$, $c_a^*(q) \geq U(q)$ for $\bar{q} \leq q \leq Q$, and $c_a^*(Q) = c_a$ for $q \geq Q$ will do. Actually the optimal policies will be perturbations of these policies that ensure the user strictly prefers $q \approx \bar{q}$ to any larger q. QED.

U(q) (given by Equation 5.20) is an indifference curve. It is the set of all points such that when the user makes purchases of size q, the frequency of purchase f is related to q by the necessary conditions for optimality, and the expected punishment for possessing q is U(q), then the user's utility is the same as it is at the best possible point under a maximum punishment policy (the utility $z(F,Q,c_a(q) = c_a)$ given by Equation 5.10B). In other words, the user is equally happy at any point along the curve U(q).

Figure 5.3 shows the general shape of U(q). It is positively sloped for $q < Q$, concave, reaches a maximum of c_a at $q = Q$, and equals zero for $q = Q + \frac{p\,c_a}{2\,c_p} \pm \frac{\sqrt{p\,c_a}}{2\,c_p}\sqrt{p\,c_a + \frac{\alpha}{h}\left(\alpha - 2\sqrt{h\left(c_S + p\,c_a\right)}\right)}$. The smaller of these two zeros is $\bar{q}$ and $0 < \bar{q} < Q$.

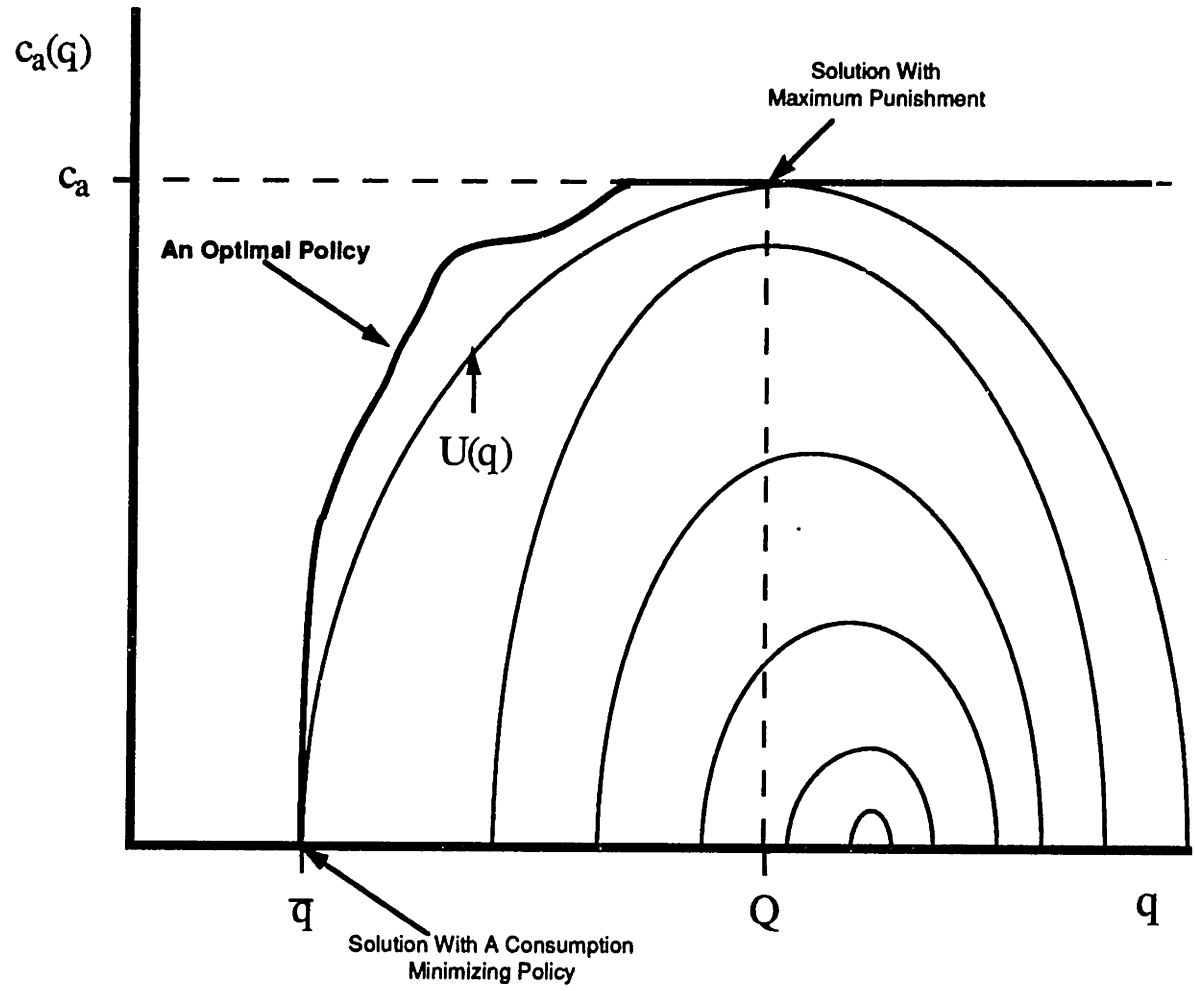
Figure 5.3:
The Indifference Curve U(q)



The theorem states that any punishment policy $c_a(q)$ that is (1) zero for quantities less than $\bar{q}$, (2) is at least as great as U(q) for $\bar{q} \leq q \leq Q$, and (3) equals c_a for $q \geq Q$ minimizes consumption. Note the consumption-minimizing policy is not unique because the value for $\bar{q} < q < Q$ is not uniquely specified.

Figure 5.4 shows one such policy. (The curves beneath U(q) are also indifference curves; lower curves represent higher utilities for the user.) It coaxes the user to reduce q from Q to $\bar{q}$ by reducing the punishment for smaller quantities. Since the rate of consumption qf is increasing in q, this also reduces consumption. As long as the punishment $c_a(q)$ must be nonnegative though, there is a limit to how much one can reduce consumption because $U(q) < 0$ for $q < \bar{q}$.

Figure 5.4:
Indifferences Curves and A Consumption Minimizing Policy



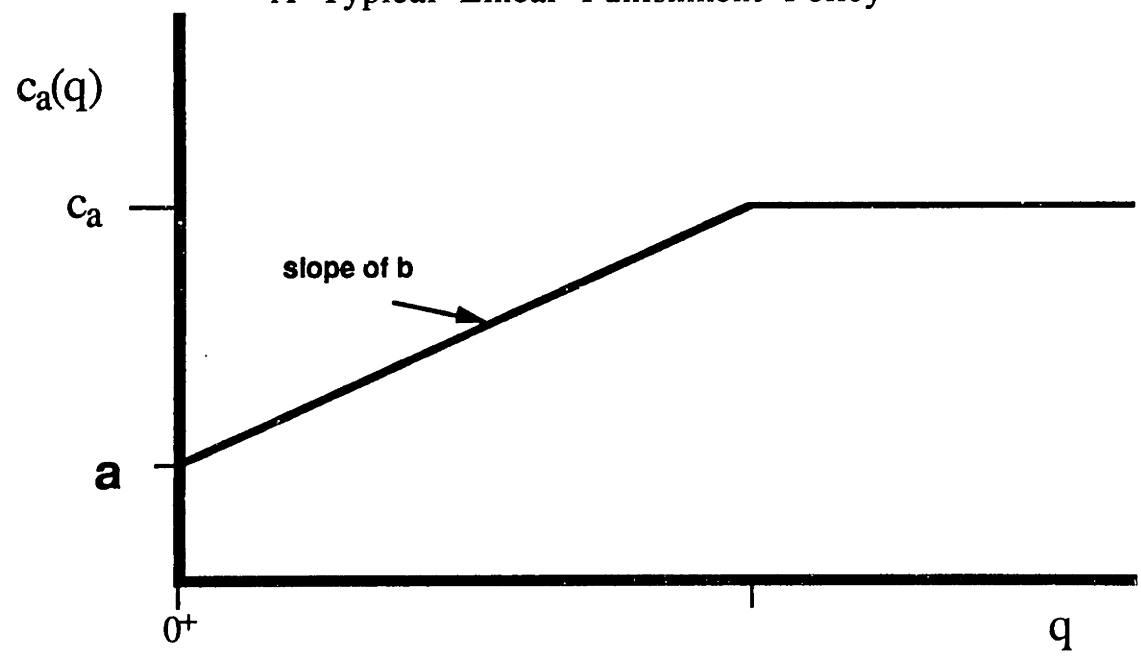
It might appear that the consumption-minimizing policy is too complicated to implement, but there is a class of policies that minimize consumption, and not all of them are complicated. For example, imposing no penalty for $q \leq \bar{q}$ and the maximum penalty c_a for $q > \bar{q}$ minimizes consumption. Nevertheless, it may be that none of these are politically feasible because they all impose no punishment for possession of quantities up to the threshold quantity $\bar{q} > 0$.

5.3.4 Consumption-Minimizing Linear Policy

It has been shown that a maximum punishment policy does not minimize consumption. The previous section described the policies that do minimize consumption, but noted that they may not be politically feasible because they impose no punishment for quantities less than the threshold $\bar{q}$. This section finds the consumption-minimizing policy within a restricted class of policies that might realistically be implemented.

One can only guess at the kinds of punishment policies that are politically acceptable, but given public sentiment at the moment, one restriction might be to consider only policies that assign some punishment for possession of any positive amount, no matter how small.[18] Also, legislators might quite reasonably object if the punishment were calculated by some complicated mathematical expression that the public could not understand. This suggests restricting attention to linear punishment policies. Then legislators would just have to specify the minimum punishment, the maximum punishment, and the smallest quantity that merits the maximum punishment, and then draw a straight line connecting these points. Figure 5.5 shows a typical linear punishment policy.

Figure 5.5:
A Typical Linear Punishment Policy



[18]This sentiment is reflected in the words of the recently released *National Drug Control Strategy*, "we need a national enforcement strategy that casts a wide net and seeks to ensure that all drug use -- whatever its scale -- faces the risk of criminal sanction." (The White House, September, 1989, p.18).

Mathematically the problem is the same as the one in the previous section except the punishment policy must be selected from a smaller set. Specifically, it is

$$\underset{c_a(q) \in G}{\text{Min}} \left\{ f^* q^* \mid (f^*, q^*) = \arg \underset{f,q \geq 0}{\text{Max}} \left\{ z(f,q) = \alpha\sqrt{fq} - hq - (c_s + p\, c_a(q))\, f - c_p\, f\, q \right\} \right\} \quad \text{(P 3)}$$

$$G = \left\{ g: \Re \rightarrow \Re \mid g(0) = 0,\ g(q) = a + bq \text{ with } a,b \geq 0 \text{ for } 0 < q \leq \frac{c_a - a}{b},\ g(q) = c_a \text{ for } q \geq \frac{c_a - a}{b} \right\}$$

Theorem 2 gives the solution. The superscript ^ is used to denote the optimal quantities for this problem. Again, the proof is rather technical and can be skipped.

Theorem 2. If

$$\alpha \leq 2\sqrt{h(c_s + p\, c_a)} \qquad (5.6D)$$

then setting $c_a(q) = c_a$ gives $\hat{q} = \hat{f} = 0$. If $\alpha > 2\sqrt{h(c_s + p\, c_a)}$ the consumption-minimizing policy has an intercept of

$$a = 0 \qquad (5.21)$$

and slope

$$b = \frac{c_p}{p} \frac{\alpha\left(\sqrt{h(c_s + p\, c_a)} - \sqrt{h\, c_s}\right) - h\, p\, c_a}{\left(\frac{\alpha}{2} - \sqrt{h(c_s + p\, c_a)}\right)^2} \qquad (5.22)$$

which gives

$$\hat{q} = \frac{\left(\frac{\alpha}{2} - \sqrt{h(c_s + p\, c_a)}\right)^2}{\left(\frac{\alpha}{2} - \sqrt{h\, c_s}\right) c_p} \sqrt{\frac{c_s}{h}}, \qquad (5.7D)$$

$$\hat{f} = \frac{\left(\frac{\alpha}{2} - \sqrt{h(c_s + p\,c_a)}\right)^2}{\left(\frac{\alpha}{2} - \sqrt{h\,c_s}\right)c_p}\sqrt{\frac{h}{c_s}}\,, \tag{5.8D}$$

$$\hat{q}\hat{f} = \left(\frac{\left(\frac{\alpha}{2} - \sqrt{h(c_s + p\,c_a)}\right)^2}{\left(\frac{\alpha}{2} - \sqrt{h\,c_s}\right)c_p}\right)^2, \text{ and} \tag{5.9D}$$

$$z(\hat{f}, \hat{q};\, c_a(q) = b^*q) = \frac{\left(\frac{\alpha}{2} - \sqrt{h(c_s + p\,c_a)}\right)^2}{c_p}. \tag{5.10D}$$

Proof: If $\alpha \le 2\sqrt{h(c_s + p\,c_a)}$ then setting $c_a(q) = c_a$ for all $q > 0$ is optimal because it yields the trivial solution.

Suppose $\alpha > 2\sqrt{h(c_s + p\,c_a)}$. It is difficult to solve over G directly, but if the feasible set G is replaced by

$$G_2 = \{g: \Re \to \Re \mid g(0) = 0,\ g(q) = a + bq \text{ with } a,b \ge 0 \text{ for } q > 0\}$$

then Section 5.2.2 gives the solution to the inner optimization problem in closed form. This can be taken advantage of by solving the problem with $G = G_2$ subject to the additional constraint that the user's utility must be at least as great as that obtained when $c_a(q) = c_a$. For convenience, that value will be denoted by κ. Thus,

$$z(f^*, q^*, c_a(q) = a + bq) \ge z(F, Q, c_a(q) = c_a) = \frac{\left(\frac{\alpha}{2} - \sqrt{h(c_s + p\,c_a)}\right)^2}{c_p} = \kappa.$$

This constraint eliminates all solutions admitted when $G = G_2$ that should not be allowed since, if the solution has $a + bq^* > c_a$, then $z(f^*, q^*, c_a(q) = a + bq) \le z(f^*, q^*, c_a(q) = c_a) \le z(F, Q, c_a(q) = c_a)$. Hence the problem reduces to

$$\min_{a,b \ge 0} \left(\frac{\left(\frac{\alpha}{2} - \sqrt{h(c_s + p\,a)}\right)^2}{c_p + p\,b}\right)$$

$$\text{subject to} \quad g(a,b) = \kappa - \frac{\left(\frac{\alpha}{2} - \sqrt{h(c_s + p\,a)}\right)^2}{c_p + p\,b} \leq 0.$$

If $\alpha > 2\sqrt{h(c_s + p\,a)}$ the objective function is decreasing in both a and b and g(a,b) is increasing. This means the solution will satisfy the inequality as an equality and thus the problem reduces to

$$\underset{a,b \geq 0}{\text{Min}} \quad \frac{\kappa}{c_p + p\,b}$$

$$\text{subject to} \quad b(a) = \frac{c_p}{p} \frac{\alpha\left(\sqrt{h(c_s + p\,c_a)} - \sqrt{h(c_s + p\,a)}\right) + h\,p\,(a - c_a)}{\left(\frac{\alpha}{2} - \sqrt{h(c_s + p\,c_a)}\right)^2}.$$

Since the objective function is decreasing in b and $\frac{db(a)}{da} < 0$ when $\alpha > 2\sqrt{h(c_s + p\,a)}$, the optimal linear policy is to choose

$a^* = 0$ and

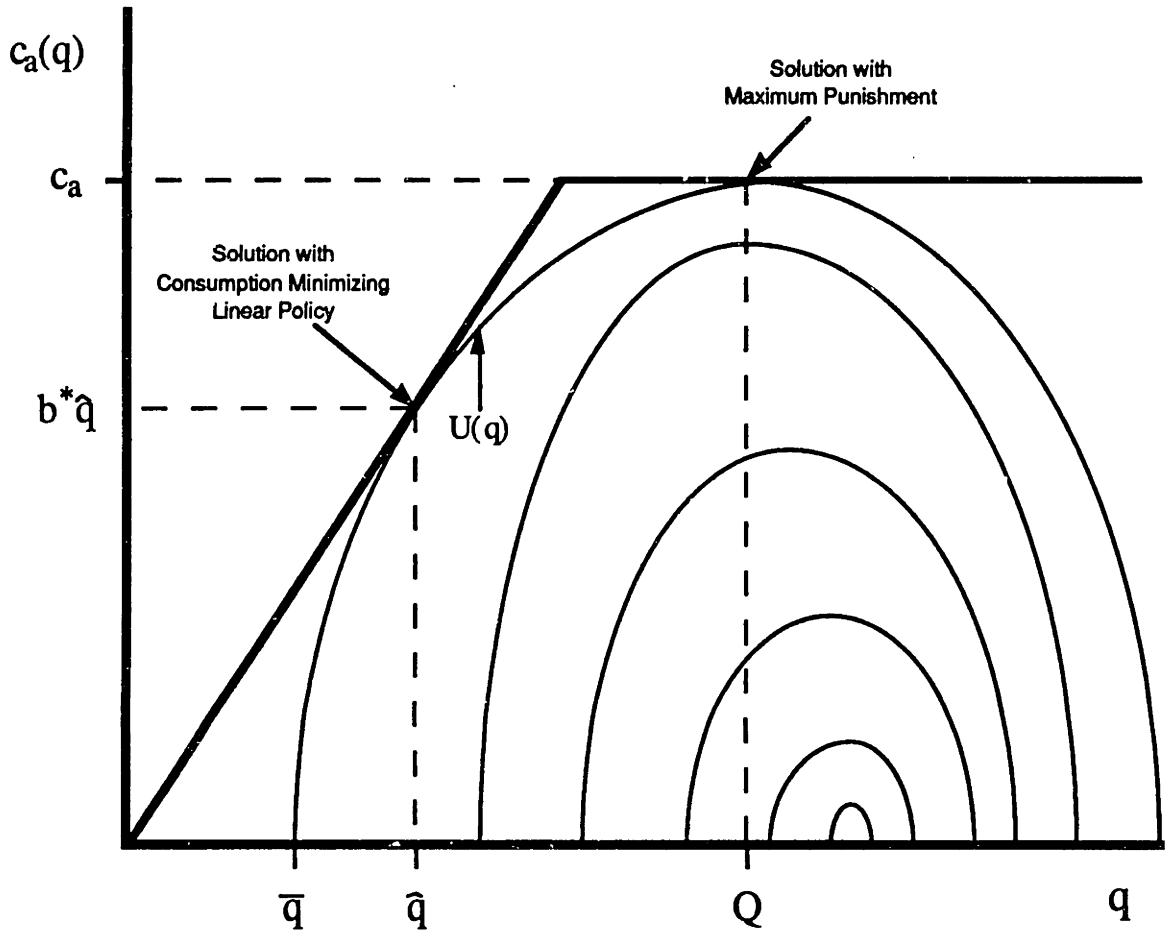
$$b^* = \frac{c_p}{p} \frac{\alpha\left(\sqrt{h(c_s + p\,c_a)} - \sqrt{h\,c_s}\right) - h\,p\,c_a}{\left(\frac{\alpha}{2} - \sqrt{h(c_s + p\,c_a)}\right)^2} \geq 0.$$

Note, $b^* > c_a/Q$, so the optimal linear policy has a steeper slope than the proportional policy considered above, and thus $a^* + b^*Q = b^*Q > c_a$. (See Figure 5.6.)

Substituting these values of a* and b* into the expressions in Section 5.2.2 gives the desired result. QED.

The expression for the optimal slope is complicated, but some algebraic manipulation shows it is greater than c_a/Q, and so it is steeper than the slope of the linear policy considered in Section 5.3.2.. (See Figure 5.6.)

Figure 5.6:
The Consumption Minimizing Linear Policy



The consumption-minimizing linear policy's intercept is zero. It imposes the lightest possible sanctions for possession of trace amounts. This is consistent with the previous section's finding that the overall consumption-minimizing policy calls for no punishment whatsoever for amounts up to some positive threshold.

Looking at Figure 5.6 shows why this is so. The policy maker must pick a straight line connecting the vertical axis with the horizontal line $c_a(q) = c_a$. The new equilibrium point will be at the place where this line is tangent to an indifference curve, either U(q) or one below U(q). The new equilibrium point will be farthest to the left (yielding the lowest consumption) if the line is tangent to U(q) itself and the slope is as steep as possible. This is accomplished by making the line's vertical intercept zero.

5.3.5 Policy of Not Punishing Users

Before comparing the policies discussed above, it is useful to consider what would happen if users were not punished, i.e. if $c_a(q) = 0$ for all q. Note, this is not a model of legalization because the search cost and price would almost certainly fall substantially if the drug were made legal. It is closer to a "decriminalization" policy that leaves antidrug laws on the books, but does not enforce them against users.

The solution can be obtained directly from the expression in Section 5.2.2 because this is the special case of Problem P1 with a = b = 0. Italics are used to denote optimal values when there is no punishment. If

$$\alpha \leq 2\sqrt{h\,c_s} \tag{5.6E}$$

then $q = f = 0$, but if $\alpha > 2\sqrt{h\,c_s}$ the solution is

$$q = \left[\frac{\alpha}{2}\sqrt{\frac{c_s}{h}} - c_s\right]\frac{1}{c_p}, \tag{5.7E}$$

$$f = \left[\frac{\alpha}{2}\sqrt{\frac{h}{c_s}} - h\right]\frac{1}{c_p}, \tag{5.8E}$$

$$qf = \left(\frac{\frac{\alpha}{2} - \sqrt{h\,c_s}}{c_p}\right)^2, \text{ and} \tag{5.9E}$$

$$z(f, q;\, c_a(q) = 0) = \frac{\left(\frac{\alpha}{2} - \sqrt{h\,c_s}\right)^2}{c_p}. \tag{5.10E}$$

5.4 Discussion

It is assumed here that the objective is to minimize consumption, so in this section the adjective "optimal" is used in place of the term "consumption-minimizing."

5.4.1 Comparison of Policies

The five policies considered above are compared below. Table 5.1 reviews the notation used for each policy.

Table 5.1:
Summary of Notation Used for Different Policies

Capital	Q	Section 5.3.1: Maximum Punishment	$c_a(q) = c_a$	
Tilda	$\tilde{q}$	Section 5.3.2: First Linear Policy	$c_a(q) = \frac{c_a}{Q} q$	$0 \le q \le Q$
			$c_a(q) \ge c_a$	$q \ge Q$
Bar	$\bar{q}$	Section 5.3.3: Optimal Policy	$c_a(q) = 0$	$0 \le q \le \bar{q}$
			$c_a(q) = c_a$	$q \ge \bar{q}$
Hat	$\hat{q}$	Section 5.3.4: Optimal Linear Policy	$c_a(q) = b^* q$	$0 \le q \le \frac{c_a}{b^*}$
			$c_a(q) = c_a$	$q \ge \frac{c_a}{b^*}$
Italics	q	Section 5.3.5: No Punishment	$c_a(q) = 0$	

If $\alpha \le 2\sqrt{h(c_s + p\,c_a)}$ the optimal linear policy, overall optimal policy, and maximum punishment policy all give the same solution, so for purposes of comparison, it is assumed that $\alpha \ge 2\sqrt{h(c_s + p\,c_a)}$.

Some tedious algebra allows one to rank the policies in terms of their ability to reduce consumption.

1) Optimal Policy
2) Optimal Linear Policy
3) First Linear Policy
4) Maximum Punishment
5) No Punishment

The ranking for the purchase size, q, and the frequency of purchase, f, depends on the parameter values. Specifically, if $\alpha >$

$2\sqrt{h}\left(\sqrt{c_s + p\,c_a} + \sqrt{c_s}\right)$ then in order of increasing purchase size the ranking is

1) Optimal Policy
2) Optimal Linear Policy
3) First Linear Policy
4) Maximum Punishment
5) No Punishment

and in order of increasing purchase frequency f it is

1) Maximum Punishment
2) Optimal Linear Policy
3) First Linear Policy
4) No Punishment.

If $\alpha < 2\sqrt{h}\left(\sqrt{c_s + p\,c_a} + \sqrt{c_s}\right)$, however, then in order of increasing purchase size the ranking is

1) Optimal Policy
2) Optimal Linear Policy
3) First Linear Policy
4) No Punishment
5) Maximum Punishment

and order of increasing purchase frequency it is

1) Optimal Linear Policy
2) Maximum Punishment
3) First Linear Policy
4) No Punishment

Note, the purchase frequency under an optimal policy depends on the specific policy and parameter values, so it cannot be ranked.

The user's utility at equilibrium can be similarly ranked in decreasing order:

1) No Punishment
2) First Linear Policy
3) Maximum Punishment (3 way tie)
   Optimal Policy
   Optimal Linear Policy.

Figures 5.7 and 5.8 show these comparisons graphically. When two quantities appear to be equal in the figures it indicates that their relative size depends on the parameter values.

Figure 5.7:
Comparison of Five Punishment Policies

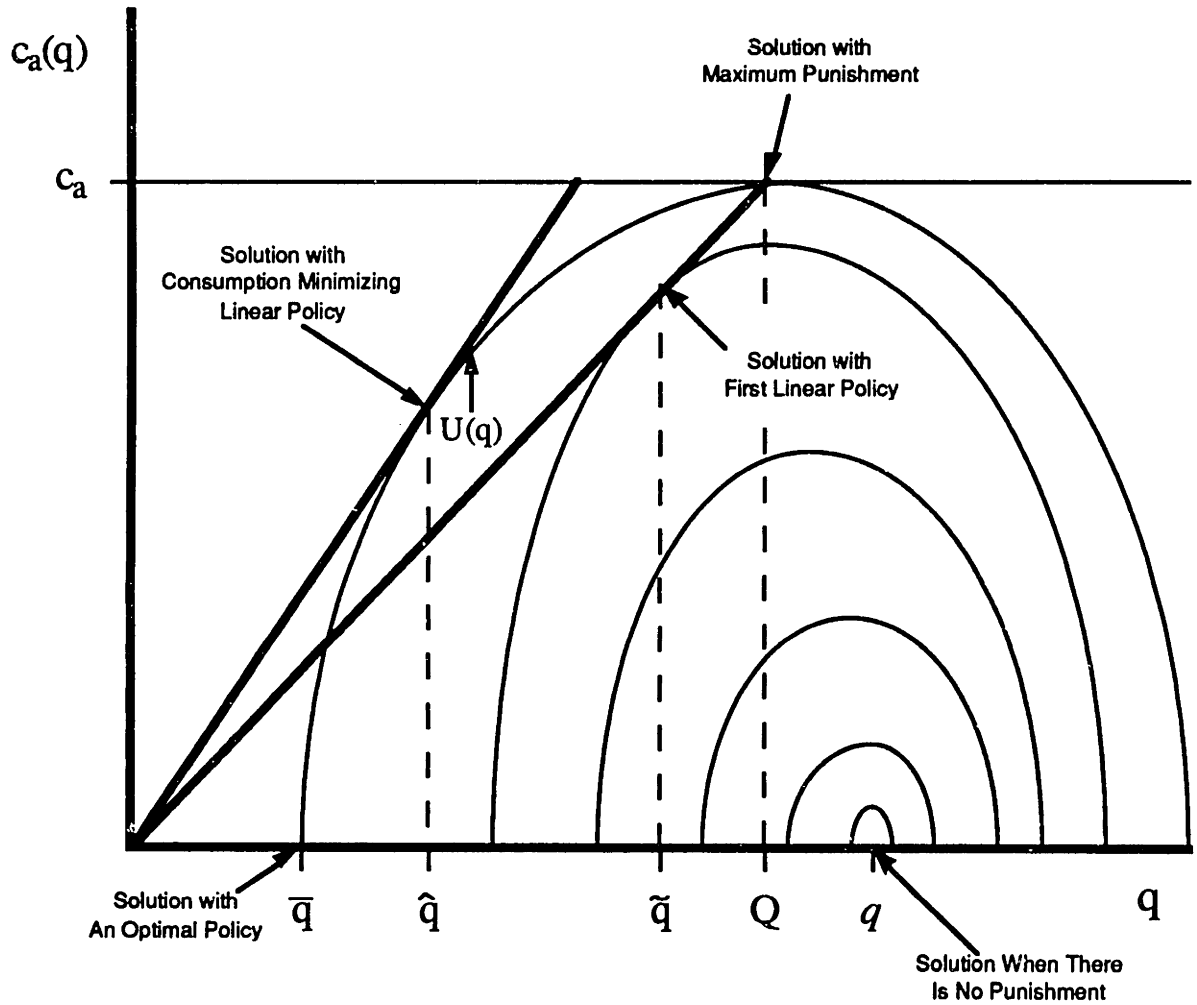
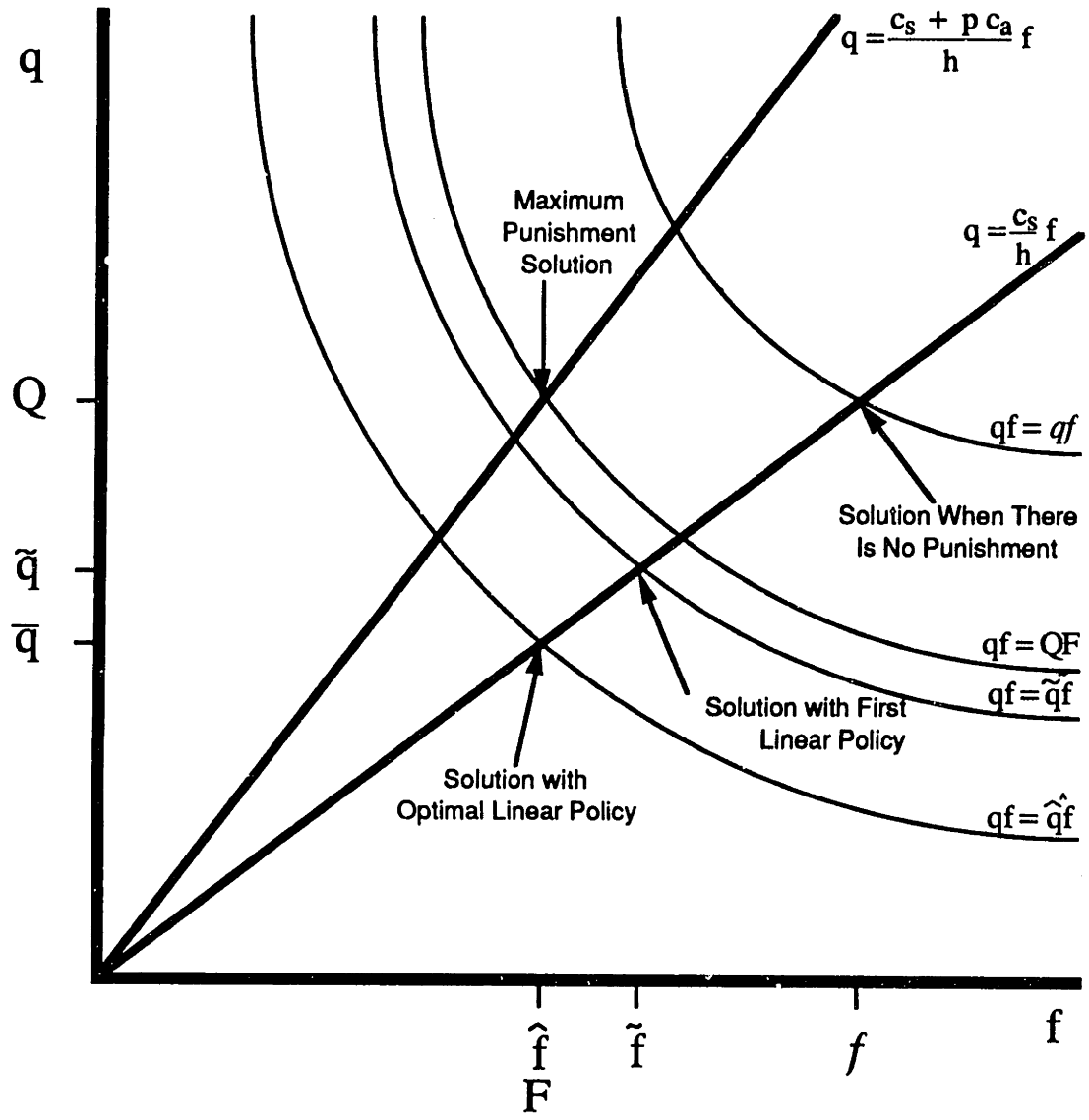


Figure 5.8:
Comparison of Policies in the q-f Plane



5.4.2 Potential for Reducing Consumption by Changing Policy

Simply signing quantities is not always very helpful. For example, Figure 5.8 shows that switching from a maximum punishment policy to the optimal linear policy will reduce consumption, but by how much? A 50% reduction might justify overcoming political and bureaucratic obstacles that a 5% reduction would not.

In general it is difficult to do more than determine relative magnitudes because the parameter values are not known, but some algebra shows that

$$\frac{\hat{qf}}{QF} = \frac{QF}{qf}. \qquad (5.23)$$

That is, the ratio of consumption under the optimal linear policy to that under a maximum punishment policy is the same as the ratio of consumption under maximum punishment to consumption with no punishment. Thus, one can (theoretically) achieve the same percentage reduction in consumption going from a maximum punishment policy to a proportional punishment policy as one can going from no punishment to maximum punishment. For example, if one believed that the current policy is essentially a maximum punishment policy and that consumption would double if users were no longer threatened with punishment, then this result suggests that theoretically one could hope to reduce current consumption by 50% by switching to a proportional punishment policy.

In many respects the United States currently has a maximum punishment policy. While penalties may increase with quantity in a staircase fashion, generally the upper limit of the lowest punishment category is more than is usually held for personal consumption.[19] Thus, the punishment facing an individual user is essentially independent of quantity.[20] There may be some quantity dependence because larger quantities encourage police to collect evidence more carefully and district attorneys to prosecute more vigorously, but this effect is probably relatively minor. In practice, punishment also depends on a variety of other factors including the offender's prior record and possibly how crowded the criminal justice system is. So the expected punishment is not a fixed, known constant, but it does not depend greatly on quantity, and that is what is important for the argument here.

Most people concede that there is at least a reasonable chance that consumption would increase appreciably if users were not threatened with punishment. This chapter suggests that if that is the

---

[19]Controlled Substances Act, Chapter 13, Subchapter I, Part D, Section 841.

[20]Thompson (1989, pp.31-32) gives examples when stiff mandatory penalties have been imposed on people who might have been treated more leniently until recently.

case, one could hope for further appreciable reductions by modifying laws to make punishment proportional to the quantity possessed. Such a reform would be in keeping with the national drug strategy which states that "Punishment should be flexible -- let the penalty fit the nature of the crime."[21]

There are, however, at least two significant qualifications to this conclusion. They are discussed in the next section.

5.4.3 Qualifications to the Conclusion Above

The discussion above is based on the observation that

$$\frac{\hat{qf}}{QF} = \frac{QF}{qf}. \qquad (5.23)$$

However, the fact that the ratios are exactly equal depends on the assumption that the satisfaction derived from using increases as the square root of the quantity consumed, and clearly that is somewhat of an arbitrary assumption. One would expect the function to be concave because of diminishing returns, but its exact form is unknown and unlikely to be so simple. The next section suggests, however, that the basic conclusion is more robust than the exact equality of the ratios.

The second and more serious qualification is that the slope of the optimal linear policy depends on the individual user's utility parameters. It is difficult to imagine how one might measure the utility parameters accurately enough to quantitatively compute the optimal slope. Furthermore, the optimal slope varies from individual to individual.

The problem with not knowing the users' utility parameters is most easily illustrated by considering an attempt to implement an optimal policy of no punishment when $q \leq \bar{q}$ and maximum punishment if $q > \bar{q}$. Since $\bar{q}$ is unknown, the policy would actually be

$$c_a(q) = \begin{cases} 0 & 0 \leq q \leq x \\ c_a & q \geq x \end{cases}$$

where x is some estimate of $\bar{q}$. Figure 5.9 shows how the resulting purchase size depends on x. If $x < \bar{q}$ (the government underestimates $\bar{q}$), the purchase size and rate of consumption will be the same as

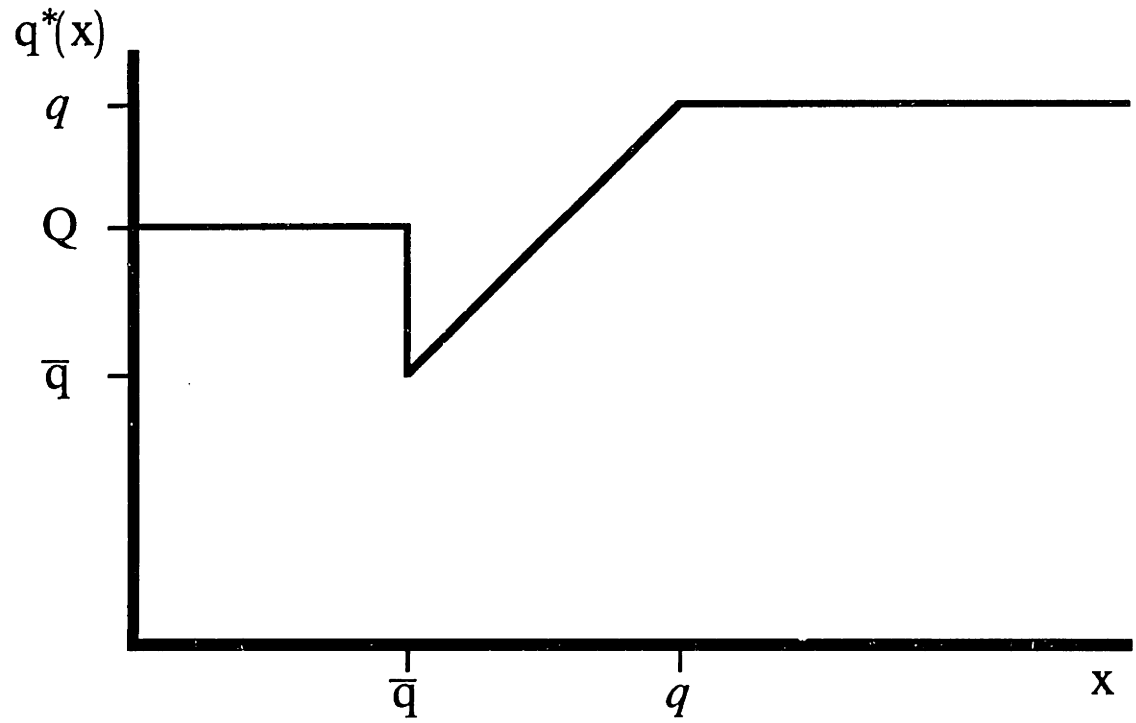
[21]The White House, 1989, p.19.

they would be under a policy of maximum punishment. If $x = \bar{q}$ the new policy will work as desired. As x increases beyond $\bar{q}$, (the government overestimates $\bar{q}$), the purchase size increases linearly. Solving for the optimal frequency, one finds that for $\bar{q} < x < q$,

$$qf = \left(\frac{\alpha x}{2(c_s + c_p x)}\right)^2, \tag{5.24}$$

so the rate of consumption rises with x until $x = q$. At that point, consumption levels off at qf, the rate of consumption when there is no punishment. That rate may be significantly higher than the original rate QF.

Figure 5.9:
Purchase Size As a Function of Permitted Quantity



Hence, if the government overestimates $\bar{q}$, changing the punishment policy will have no effect on consumption, and if it underestimates $\bar{q}$ enough, consumption might actually increase!

Attempting to implement the optimal linear punishment policy poses a similar dilemma. If the slope is too steep, changing from a policy of maximum punishment will have no effect, but if the slope is

too shallow, changing the punishment policy may increase rather than decrease consumption.

5.5 Generalizations

This section seeks to generalize the results above. The first subsection relaxes the assumption that the benefit of using be proportional to the square root of the quantity consumed. The second subsection discusses implications of applying the model to dealers as well as users. The third subsection briefly examines the possibility of making punishment depend on the number of previous arrests as well as the quantity possessed at the time of arrest.

5.5.1 Generalizing the Form of the Benefits of Using Term

A valid criticism of the foregoing analysis is that it assumes a very specific functional form for the benefit of consuming, $B(f,q) = \alpha \sqrt{qf}$. This section shows that the key results can be obtained without making this assumption.

It was argued above, and will still be assumed here, that B(f,q) is a function only of the rate of consumption (fq), not of q or f separately, and that as a function of qf it is increasing, concave, and zero when there is no consumption, i.e. that

$$\begin{aligned} &B(f,q) = B(fq), \\ &B(0) = 0, \\ &B'(x) > 0, \text{ and} \\ &B''(x) < 0. \end{aligned} \tag{5.25}$$

It will further be assumed that B(x) is continuously differentiable.

Assuming a linear punishment policy, the user's problem is to

$$\begin{aligned} &\text{Max } z(f,q) = B(fq) - hq - (c_s + pa)\, f - (c_p + pb)\, qf \\ &\text{s.t.} \quad q,f \geq 0. \end{aligned} \tag{P4}$$

The first order conditions are

$$\nabla z(f,q) = \begin{bmatrix} \dfrac{\partial z(f,q)}{\partial f} \\ \dfrac{\partial z(f,q)}{\partial q} \end{bmatrix} = \begin{bmatrix} \left(B'(f^*q^*) - (c_p + pb)\right) q^* - (c_s + pa) \\ \left(B'(f^*q^*) - (c_p + pb)\right) f^* - h \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \tag{5.26}$$

They imply that

$$q^* = \frac{c_s + pa}{h} f^* \tag{5.27}$$

as before. Hence consumption is

$$q^* f^* = \frac{h}{c_s + pa} q^{*2}. \tag{5.28}$$

Let u denote this quantity.

The user's problem can be reformulated as a one dimensional optimization problem in u, the rate of consumption.

$$\underset{u \geq 0}{MAX}\, Z(u) = B(u) - 2\sqrt{h(c_s + pa)}\sqrt{u} - (c_p + pb)u. \tag{P5}$$

The first order condition is

$$Z'(u^*) = 0 = B'(u^*) - \frac{\sqrt{h(c_s + pa)}}{\sqrt{u^*}} - (c_p + pb), \tag{5.29}$$

and the second order necessary condition for a maximum is

$$Z''(u^*) = B''(u^*) + \frac{\sqrt{h(c_s + pa)}}{2\, u^{*3/2}} \leq 0. \tag{5.30}$$

The implicit function theorem can be applied to the first order Condition 5.29 to determine how consumption varies as a and b change. Specifically, let F(a,b,u) = Z'(u). Suppose that u_0 is the solution for two particular parameter values $a = a_0$ and $b = b_0$.

If the second derivative is strictly negative at the optimum (i.e. Condition 5.30 holds with a strict inequality), then F_u is nonsingular, so the implicit function theorem states that there is a neighborhood M of (a_0, b_0, u_0) in $\Re^3$, a neighborhood N around (a_0, b_0), and a function $U^*: \Re^2 \dashrightarrow \Re^1$ such that

$U^*(a_0,b_0) = u_0$,
$(a,b,U^*(a,b)) \in M$ for $(a,b) \in N$, and
$F(a,b,U^*(a,b)) = 0$ for $(a,b) \in N$.

Moreover, U* is continuously differentiable and

$$\frac{dU^*(a,b)}{da} = \frac{-F_a(a,b,G(a,b))}{F_u(a,b,G(a,b))} = \frac{\dfrac{p\sqrt{h}}{2\sqrt{(c_s+pa)\,G(a,b)}}}{B''(G(a,b)) + \dfrac{\sqrt{h(c_s+pa)}}{2\,G(a,b)^{3/2}}} \qquad (5.31)$$

$$\frac{dU^*(a,b)}{db} = \frac{-F_b(a,b,G(a,b))}{F_u(a,b,G(a,b))} = \frac{p}{B''(G(a,b)) + \dfrac{\sqrt{h(c_s+pa)}}{2\,G(a,b)^{3/2}}}.$$

Equation 5.30 implies the denominator is negative, so both derivatives are negative. That is, the optimal rate of consumption is decreasing in both a and b in a neighborhood around the solution. This is to be expected. The more interesting question is, what happens if parameter a decreases while b increases, as would be the case if a maximum punishment policy were converted into a proportional punishment policy?

Suppose that for all q greater than or equal to some pivot quantity Q_p the punishment will be the same as before. The punishment policy for $q < Q_p$ must still be linear, but it can have a different slope. Then

$$a + b\,Q_p = a_0 + b_0\,Q_p, \qquad (5.32)$$

so as b increases, a would have to decrease by Q_p times as much.

If Q_p were less than the purchase size corresponding to the consumption rate u_0 ($\sqrt{\frac{h}{c_s+pa}}\sqrt{u_0}$), then incremental changes in b would have no effect. So suppose $Q_p > \sqrt{\frac{h}{c_s+pa}}\sqrt{u_0}$.

Now the user's problem can be reformulated as

$$\underset{u\geq 0}{MAX}\, Z(u) = B(u) - 2\sqrt{h(c_s+p[a_0+(b_0-b)Q_p])}\sqrt{u} - (c_p+pb)u. \qquad (P5b)$$

The first order condition is

$$Z'(u^*) = B'(u^*) - \frac{\sqrt{h\left(c_s + p\left[a_0 + (b_0 - b)\,Q_p\right]\right)}}{\sqrt{u^*}} - (c_p + pb) = 0, \tag{5.29b}$$

and the second order necessary condition for a maximum is

$$Z''(u^*) = B''(u^*) + \frac{\sqrt{h\left(c_s + p\left[a_0 + (b_0 - b)\,Q_p\right]\right)}}{2\,u^{*3/2}} \leq 0. \tag{5.30b}$$

Assuming this second order condition is satisfied as a strict inequality, the implicit function theorem can be applied to F(b,u) = Z'(u) around the point (b_0,u_0) to obtain

$$\frac{dU^*(b)}{db} = \frac{-F_b(b,G(b))}{F_u(b,G(b))}$$

$$= \frac{p - \sqrt{\frac{h}{G(b)}}\,\frac{p\,Q_p}{2\sqrt{c_s + p\left[a_0 + (b_0 - b)\,Q_p\right]}}}{B''(G(b)) + \frac{\sqrt{h\left(c_s + p\left[a_0 + (b_0 - b)\,Q_p\right]\right)}}{2\,G(b)^{3/2}}}. \tag{5.31b}$$

The second order condition implies the denominator is negative. So the rate of consumption is decreasing in b around the point (b_0,u_0) if and only if

$$p > \sqrt{\frac{h}{u_0}}\,\frac{p\,Q_p}{2\sqrt{c_s + p\,a_0}}, \tag{5.32}$$

i.e. if and only if

$$Q_p < 2\sqrt{\frac{c_s + p\,a_0}{h}}\,\sqrt{u_0} = 2\,q_0, \tag{5.33}$$

where q_0 is the purchase quantity before the policy change.

So the quantity consumed will decrease when the slope of the punishment policy increases as long as the pivot point Q_p is less than

twice the current purchase size. Looking back at Equation 5.31, this result makes sense. It states that

$$\frac{dU^*(a,b)}{db} = 2\,q_0 \frac{dU^*(a,b)}{da}. \quad (5.34)$$

Furthermore, the larger Q_p is, the larger F_b is, and hence the larger (less negative) the change in consumption is. So reforming the punishment policy will have the greatest effect if Q_p is only slightly larger than q_0.

This result is encouraging because it shows that the main point derived above, that changing policy from maximum punishment to proportional punishment may reduce consumption, holds more generally. Also, it shows how accurately one must estimate the current purchase quantity to avoid having consumption increase rather than decrease when policy is reformed to make punishment increasing in quantity.

The second and more fundamental problem discussed in Section 5.4.3 remains. The best policy depends on the user's parameters. These parameters are hard to measure and vary from user to user.

5.5.2 Generalizing the Interpretation to Dealers

Subsection 5.5.1 generalized the foregoing results mathematically. The benefit function $B(f,q) = \alpha \sqrt{qf}$ was replaced by any function satisfying Conditions 5.25. The results can also be generalized simply by applying them to dealers as well as to users.

The model fits anyone who

(1) regularly buys some quantity q with some frequency f,
(2) incurs search time costs proportional to their purchase frequency,
(3) incurs holding costs proportional to the purchase size q,
(4) pays a constant dollar price (no quantity discounts),
(5) is arrested with probability proportional to the frequency of purchase,
(6) receives punishment upon arrest that is linear in the purchase quantity, and
(7) derives benefits from obtaining (e.g. by reselling) drugs that satisfy Conditions 5.25.

There are a number of reasons why dealers might not fit these assumptions as well as casual users do. Perhaps the most serious of

these is that quantity discounts are certainly available. For modest variations in the purchase quantity, however, it may not be too grievous an approximation to ignore them.

Inventory behavior for dealers, especially high-level dealers, may also be different than it is for casual users. Dealers may contact customers before they receive a shipment so they can sell the drugs quickly after they arrive. If so, then inventory costs might not depend on q alone. If a significant portion of the inventory costs come from the risk of arrest at times when there is some inventory, then inventory costs might actually decrease in q because with smaller, more frequent purchases, there might be some drugs in inventory a greater fraction of the time. On the other hand, the temptation for others to rob or defraud the dealer is probably increasing in q.

At first one might think the benefits of dealing would be linear in the flow of drugs, not strictly concave as is assumed above. Presumably there is some heterogeneity among customers, however. Dealers sell first to the highest paying, most trustworthy customers. Hence the marginal benefit derived decreases as the dealer seeks less and less desirable customers, and the return need only diminish slightly to make the benefits concave if they would otherwise be linear.

This discussion about whether the model accurately reflects the situation facing a dealer could be extended, but the conclusion would probably remain the same. The model is imperfect; it clearly oversimplifies the true situation. Nevertheless, it does not do so in ways that are completely "wrong". Recognizing that it is an imperfect simplification, one can still ask whether it produces any useful insights.

The straightforward interpretations of the results such as

$$q^* = \frac{c_s + pa}{h} f^* \tag{5.27}$$

probably retain their meaning. For a fixed flow of drugs, if holding costs are high, dealers will tend to make many, small transactions. On the other hand, if search time costs, the probability of arrest per transaction, and/or the fixed punishment per arrest (a) are high, dealers will tend to make fewer, larger purchases.

More importantly, the basic lesson may still hold. It may be possible to change punishment policies in a way that "bribes" dealers to reduce the size of the transactions they conduct. This could increase their search time (transaction) costs, shifting up the supply

curve. Shifting up the supply curve is the basic goal of most enforcement; it increases prices and reduces the quantity consumed.

One possible drawback is that this could well increase the number of dealers, even if it reduces the total volume of sales, because the volume of business conducted by each dealer decreases. Reforming punishment policies to make them increasing in quantity would create a diseconomy of scale, giving smaller operators a better chance of competing effectively.

Whether this is good is hard to say. Having fewer, larger operators may be bad because they are more powerful, more capable of corrupting, more violent, and more efficient. On the other hand, the more dealers there are, the more people there are with careers outside the law.

A compromise may yield the "least of both bads." Ideally there would be as few dealers as possible without having any of them become very wealthy or very powerful. This could happen if drugs entered the country in relatively small quantities, and retail dealers serviced many customers.

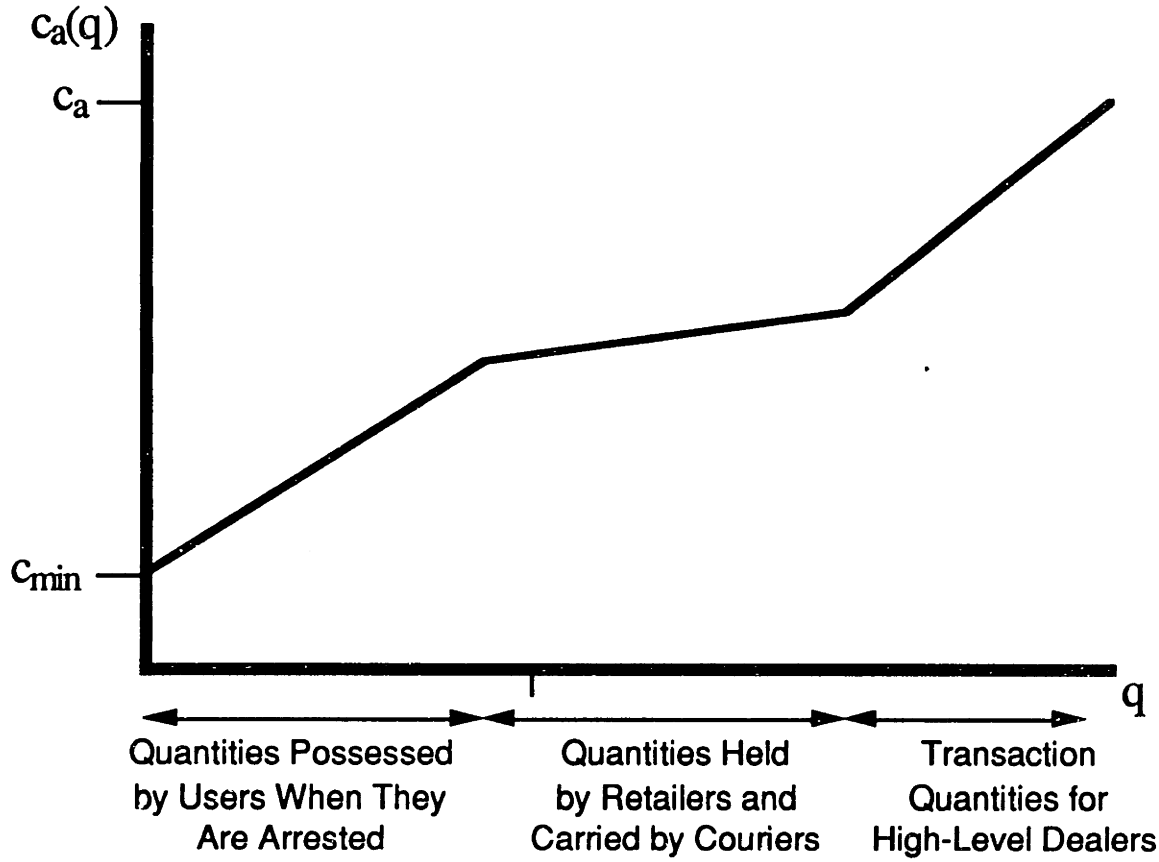
People importing a kilogram of cocaine a week might make a lot of money, but they will probably never have enough money or power to threaten this country's political institutions. That is, 2500 people importing one kilogram a week may pose less of a threat than 5 dealers importing 500 kilograms a week.

At the other end of the distribution chain, the fewer retailers there are, the fewer dealers there are in total, because the vast majority of dealers are retailers. Furthermore, reducing the number of retailers may reduce street violence, reduce the recruitment of new users and dealers, increase search time costs (especially for those who do not have established connections), and generally make it easier for local police to control the streets. That is, all other things equal, it may be less damaging to have 100 dealers selling 100 bags per day than 500 dealers selling 20 bags per day.

If this is true and the model's basic insight carries over to dealers, then it might be best to have a punishment policy such as the one depicted in Figure 5.10.

Punishment should be increasing for personal consumption quantities to encourage users to buy smaller quantities for the reasons discussed above. For large quantities punishment should be increasing to put very big dealers at a competitive disadvantage.

Figure 5.10:
A Punishment Policy for Larger Quantities



The only reason punishment should not be increasing for intermediate quantities is that "there is no free lunch." If there is some minimum acceptable punishment for trace quantities; there is some maximum possible punishment; and the punishment policy is nondecreasing, then the steeper the slope in one region, the shallower it must be somewhere else. Since strong cases can be built for why the slope should be steep for small and very large quantities, then it may be best to have a shallower slope for intermediate quantities.

5.5.3 Different Punishment Policies for Repeat Offenders

One of the limitations of the model above is that it does not explicitly keep track of users' criminal records. It considers punishment policies which are functions only of the quantity possessed at the time of arrest. Punishment can also depend on the number of arrests, $c_a(q;i)$, where i is the total number of arrests (up

to and including the most recent one). This subsection suggests how this possibility might be exploited.

The main point of this chapter is that it may be desirable to make punishment an increasing function of the quantity possessed. The extent to which it can be so is limited, however, by the minimum and maximum punishments society will allow. Having different punishment policies for repeat offenders helps circumvent this limitation in two ways. First, the minimum and maximum acceptable punishments are probably different for first-time and repeat offenders. Second, if the frequency of purchase (f) is positively related to the purchase size (q), and the probability of arrest is proportional to the frequency of purchase, then repeat offenders are more likely to be users who prefer to purchase larger amounts. That is, users' arrest histories may indirectly give information about their purchase size, and as seen above, that information is essential for choosing the proper pivot point for that class of user when maximum punishment policies are converted to proportional policies.

While the latter benefit of making punishment policies different for repeat offenders is more interesting mathematically, the first is probably more important. The upper bound on punishment severity comes from the class of users who are "least culpable." Practically speaking, the maximum punishment for personal consumption amounts is determined by the maximum punishment society will tolerate for experimental users. Frequently people arrested with small amounts are actually retail dealers. Society might be willing to punish them more severely, but it can not because the quantity possessed does not distinguish the two classes of offenders. As will be seen, however, repeated arrest for possession of such amounts may do just that. A very simple model of arrest careers helps illustrate the point.

The model makes the following assumptions.

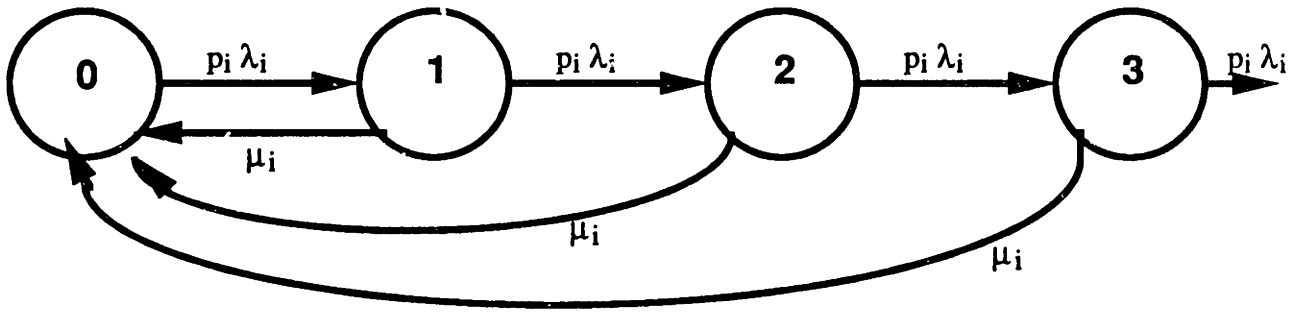
(A1) There are I distinct classes of people being arrested for possession of relatively modest quantities of drugs.
(A2) Members of each class make purchases according to a Poisson process with rate λ_i, i = 1, 2, ..., I.
(A3) There is a constant arrest risk p_i per transaction, i = 1, 2, ..., I.
(A4) The probability of arrest is independent for all transactions.
(A5) The number of people in each class is a constant N_i.
(A6) People in class i are involved with drugs for a time that is exponentially distributed with mean $1/\mu_i$.

(A7) When someone exits the population that person is replaced by someone in the same class who initially has a "clean" arrest record.
(A8) Incarceration time is negligible.

The last assumption deserves some justification. First, frequently incarceration is in fact negligible. Many people arrested for minor drug charges are released on bail and sentenced only to probation. Second, not all punishments involve incarceration. Alternative sentencing is increasingly discussed.[22] Third, the model can be readily extended to include exponentially distributed incarceration times. It makes some of the the expressions and calculations more tedious, so these details are omitted here.

An individual's arrest career can be modelled as a Markov process in which the state is the number of previous arrests. (See Figure 5.11.)

Figure 5.11:
Markov Model of User's Career



The limiting-state probability of a class i offender being in state j is

$$P_{ij} = \left(\frac{\mu_i}{\mu_i + p_i\lambda_i}\right)\left(\frac{p_i\lambda_i}{\mu_i + p_i\lambda_i}\right)^j \quad j = 0, 1, 2, \ldots \qquad (5.28)$$

which is a shifted-geometric distribution.

---

[22]See, for example, Hillsman, Mahoney, Cole, and Auchter (1987) and du Pont (1985).

For this illustration consider four classes of people: (1) experimental users, (2) casual/recreational users, (3) heavy users, and (4) retail dealers.

Experimental users as defined here are people who buy the smallest possible amount a couple of times a year for two or three years. Casual users buy more regularly, but not so often that it dominates their life. They might buy the standard retail quantity (for example, a $20, quarter-gram bag of cocaine) about once a week for five or so years. Most users are either experimental or casual users.

Heavy users buy every other day or so. As the model above suggests, they are likely to buy larger quantities than casual users. For example, they might buy "eight-balls" of cocaine (2-3 grams for $100). They are also likely to be involved with drugs for longer periods of time, perhaps ten years or so. There are not as many heavy users as there are casual users, but they still number in the millions.

It will be assumed that the arrest risk per transaction for users is about 0.0005 since there are roughly 500,000 arrests for possession each year[23] and there are very roughly about one billion retail transactions per year.[24]

Retailers obviously sell all quantities that people want to purchase. What distinguishes them from users is the frequency with which they participate in drug transactions. Over the course of a year they might average five transactions per day. It will be assumed that they are active for about ten years and their arrest risk per transaction is also 0.0005 because very roughly the same number of dealers as users are arrested,[25] and obviously every purchase is also a sale.

Table 5.2 summarizes the parameter values and limiting-state probabilities for the various classes of users. Note, these values are not in any way meant to be "real" data. They are used only for illustration. It is not even important to determine which drug is being considered or whether it is all drugs lumped together.

---

[23]See Figure 2.4.

[24]Based on estimates that on the order of $20-100 billion are spent on drugs each year.

[25]Again, see Figure 2.4.

Table 5.2:
Characteristics of Different Classes of Potential Arrestees

	Experimenters	Casual Users	Heavy Users	Dealers
N_i	12.5×10^6	12×10^6	4×10^6	0.5×10^6
λ_i (yr^{-1})	2	25	150	1850
μ_i (yr^{-1})	0.4	0.2	0.1	0.1
p_i	0.0005	0.0005	0.0005	0.0005
$p_i\lambda_i$	0.001	0.0125	0.075	0.925
P_0	0.998	0.941	0.571	0.098
P_1	0.002	0.055	0.245	0.088
P_2	0.000	0.003	0.105	0.079
P_3	0.000	0.000	0.045	0.072
P_4	0.000	0.000	0.019	0.065
P_5	0.000	0.000	0.008	0.058
P_6	0.000	0.000	0.004	0.053
P_7^+	0.000	0.000	0.003	0.487

Table 5.3 shows the expected number of arrests per year for each class broken down by whether the arrests are first arrests, second arrests, third, and so on. It clearly shows that the vast majority of times experimental users are arrested, it is their first arrest. Likewise most arrests of casual users are for first or second offenses. Hence the vast majority of people arrested three or more times are heavy users or dealers, even though those groups together make up less than one-sixth of the people who might be arrested.

Table 5.3:
Number of Arrests per Year in Each Category

	Experimenters	Casual Users	Heavy Users	Dealers
Total # of Arrests/yr.	12,500	150,000	462,500	925,000
# of 1st arrests	12,469	141,176	171,429	45,122
# of 2nd arrests	31	8,304	73,469	40,720
# of 3rd arrests		488	31,487	36,747
# of 4th arrests		29	13,494	33,162
# of 5th arrests		2	5,783	29,927
# of 6th arrests		1	2,479	27,007
# of 7th arrests			1,062	24,372
# of 8th arrests			455	21,994
# of 9th arrests			195	19,849
# of 10th arrests			84	17,912
# of 11th arrests			36	16,165
12th or more			27	149,523

Table 5.4 shows the probability that someone with a given number of arrests belongs to each of the classes. It clearly shows the potential for making punishment depend on the arrest history.

Reluctance to impose harsh sanctions for fear they will fall on relatively harmless experimenters (who will on average buy drugs no more than half a dozen times in their life) need only constrain the maximum punishment for first-time offenders. Assuming arrest probabilities for different purchases are independent, there is very little chance that an experimental user will be arrested more than once. More importantly, there is very little chance that someone arrested more than once is an experimental user.

Table 5.4:
Probability an Arrestee with a Given Number of Total Arrests Belongs to Each Category

# of Arrests	Experimenters	Casual Users	Heavy Users	Dealers
1	0.034	0.381	0.463	0.122
2	0.000	0.068	0.600	0.332
3	0.000	0.007	0.458	0.535
4	0.000	0.001	0.289	0.710
5	0.000	0.000	0.162	0.838
6	0.000	0.000	0.084	0.916
7	0.000	0.000	[illegible]42	0.958
8	0.000	0.000	0.020	0.980
9	0.000	0.000	0.010	0.990
10	0.000	0.000	0.005	0.995
11	0.000	0.000	0.002	0.998

Likewise, one might want to impose different sanctions on casual and heavy users. Presumably casual users pose less of a threat to society because they are less likely to have lost control of their lives and less likely to resort to property crime to finance their habit. Fines and community service might be deemed appropriate punishment for casual users, while compulsory drug treatment or incarceration might be more appropriate for heavy users. Table 5.4 suggests distinctions between casual and heavy users can be (partially) made based on arrest histories.

Also, in as much as heavy users probably purchase larger quantities and repeat offenders are more likely to be heavy users, it might be a good idea to have the pivot point be larger for repeat offenders. Specifically, it might be a good idea to convert a maximum punishment policy for first offenders to a proportional policy with the pivot point slightly larger than the quantity typically purchased by casual users. As discussed above, this might reduce their consumption and, assuming heavy users buy larger quantities, it would have no impact on heavy users.

For second and third offenses, however, it might be a good idea to choose a pivot point just above the quantity typically purchased by heavy users. This might help reduce their consumption and presumably would have minimal impact on experimental and casual users because they rarely face the punishment for repeat offenders.

Finally, if the person arrested has four or five previous arrests, there is a very good chance that person is not a user at all but a dealer. So punishment policies for such multiple offenders could be designed for dealers. For example, they might include mandatory jail terms.

There are limits to this approach. For one, the correlation between purchase quantity and consumption is probably actually weaker than the model developed in this chapter suggests. Second, if punishment for the first four offenses is aimed at users, then a significant fraction of dealers will never suffer from sanctions directed at dealers, and there will be an incentive for more people deal until they get their fourth arrest. Finally, this approach may not be readily generalizable to other crimes because there are relatively few crimes for which there is such heterogeneity in offenders and offense rates. Society would probably be unwilling to let people off lightly because they are "experimental murderers." Traffic and parking violations may be one of the only other infractions to which this approach could be applied.

5.6 Summary

This chapter argues that a "zero-tolerance" policy of imposing a maximum level of punishment irrespective of the quantity possessed may not minimize consumption. Switching from such a policy to one in which punishment is proportional to quantity could increase some users' welfare while simultaneously reducing consumption. In theory, consumption might fall by a fraction comparable to that which would accompany a move from no punishment to a policy of maximum punishment.

The consumption minimizing policy is derived, but it may not be politically feasible because it calls for no punishment for quantities below a certain threshold. So the consumption minimizing policy within a restricted class of policies that are more likely to be politically feasible is also derived.

However, for a variety of reasons the model is not suitable for quantitatively computing the optimal punishment policy. First of all, the model makes many assumptions that may not hold exactly. For example, it assumes that search time costs and the probability of

arrest per purchase are independent of the frequency of purchase; that may not hold exactly. It also ignores or treats only briefly important issues such as substitution between drugs, the treatment of repeat offenders, and the ability of policy makers to control parameters such as price, search time costs, and the probability of arrest. Most significantly though, the policies prescribed are functions of the user's utility parameters. These parameters cannot be measured, and even if they could be, they vary from user to user.

These points, however, do not negate the chapter's principal conclusion that a "zero-tolerance" policy for users may not minimize consumption. There are other compelling reasons for not following a "zero-tolerance" policy. It commits scarce enforcement resources to relatively minor offenders, crowds courts and prisons, and violates the principle that "the punishment should fit the crime." Nevertheless, perhaps in response to calls to "get tough" on drugs, the United States currently has what is in many respects a "zero-tolerance" policy for users. This chapter suggests that the desire to be "tough on drugs," at least in as much as that represents a desire to minimize consumption, should not preclude consideration of a policy that makes punishment proportional to the quantity possessed.

Chapter 6: AIDS' Impact on the Number of Intravenous Drug Users

6.0 Introduction

Intravenous (IV) drug users constitute the second largest subpopulation of persons with AIDS, and the rate at which new cases are diagnosed is increasing.[1] Seroprevalence levels in cities such as New York, where the epidemic is already in an advanced state, are about 60%,[2] and estimates of the ensuing health costs are staggering.

It is not surprising, therefore, that researchers have studied how the virus spreads among IV drug using populations. To date most have focused on forecasting the prevalence (fraction of all users infected) of the human immunodeficiency virus (HIV). This chapter takes a different tack. It asks how the disease will affect the size of the IV drug using population.

The simple answer is that, other factors (including drug prices, enforcement and punishment policies, availability of treatment, and so on) remaining equal, AIDS will reduce the size of the IV drug using population. This will happen because fewer potential users will start, more current users will quit, and others will die prematurely of the disease.

There are two reasons for attempting to forecast the change in the IV drug using population. The first is simply to help make population projections more accurate. The second is to provide a baseline against which one can measure the success of new programs designed to reduce the number of IV drug users.

To elaborate on the second point, suppose that in the absence of any change in policy, AIDS would cut the IV drug using population by a quarter. Suppose further that a new policy is implemented and the population decreases by one-third. It might appear that the program was a resounding success and that it should be imitated across the country, when in fact much of the program's apparent "success" in reducing the size of the IV drug using population would more accurately be attributed to the disease.

The next section introduces a simple model that describes AIDS' impact on the number of IV drug users. An expression for the

---

[1]Centers for Disease Control, April 1990. Somewhat surprisingly the rate of increase in new cases among IV drug users was lower than the overall rate of increase in new cases.

[2]Des Jarlais, 1988.

steady state population size is obtained, and factors affecting the values of the parameters appearing in that expression are discussed. Then a more detailed model of needle sharing is described. The second model is an open population version of a model Kaplan[3] developed to predict how the HIV virus would spread among an IV drug using population of fixed size. Somewhat surprisingly, both models yield essentially the same expression for the long term population change. For reasonable parameter estimates, the change is substantial, perhaps exceeding 50%.

6.1 A General Model of AIDS and IV Drug Users

The total number of IV drug users was probably relatively stable for a decade or so starting in the late 1970's. At least heroin use seems to have leveled off.[4] Other drugs such as cocaine, amphetamines, and diazepam (Valium) are injected, and there are indications that this may be increasing.[5] Nevertheless, heroin is still the most commonly injected drug, so many studies[6] use the number of heroin users as a surrogate for the number of IV drug users, as will be done here.

Three recent events have brought an end to this stability. The first is the explosion in cocaine use, particularly in the form commonly known as "crack." Since some stimulant users turn to sedatives when they "burn out," this may lead to increased heroin consumption.[7] Also, "speedballing" (consuming a mixture of cocaine and heroin) is increasingly common, as is evidenced by the rise in the number of emergency room and medical examiner mentions of cocaine and heroin in combination that are reported by the DAWN system. (See Figures 6.1 and 6.2.[8])

---

[3]Kaplan, 1989.

[4]This statement is made by U.S. Department of Justice (1989), Marshall (1988a), Marshall (1988c), and Isikoff (1989a). Also, between 1979 and 1987 the fraction of seniors interviewed for The High School Senior Survey who reported having ever used heroin, used it within the last year, and used it within the last month were always within 0.1% of 1.2%, 0.6%, and 0.3% respectively (U.S. Department of Health and Human Services, 1988c.)

[5]Committee on AIDS Research, 1989, p.196.

[6]E.g. Kleiman and Mockler, 1988.

[7]*Time* (1990) reports that this is in fact happening as "an increasing number of cocaine abusers are using heroin to ease the horror of the postcrack low."

[8]Sources: 1976-1985 data, U.S. Dept.of Health and Human Services, 1987b, p.67 and p.85; 1987 data, U.S. Dept. of Health and Human Services, 1988d, p.44 and p.67; and 1988 data, U S. Dept. of Health and Human Services, 1989d, p.44 and p.67.

Figure 6.1
DAWN Reported Emergency Room Mentions for Cocaine and Heroin in Combination

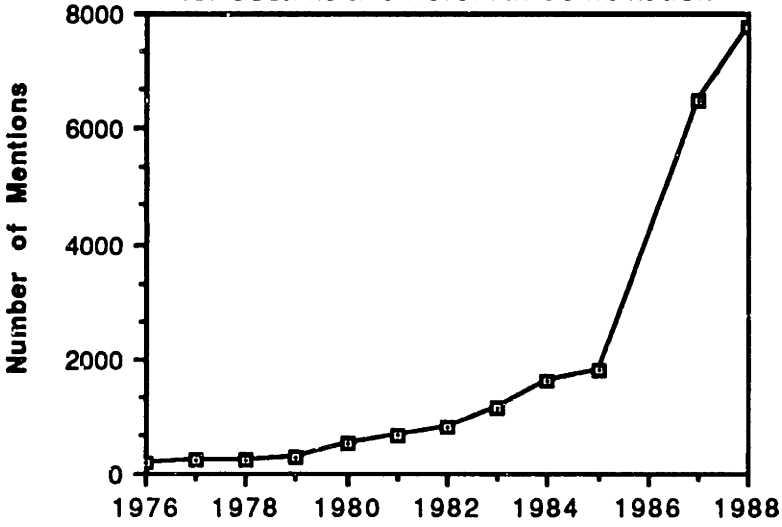
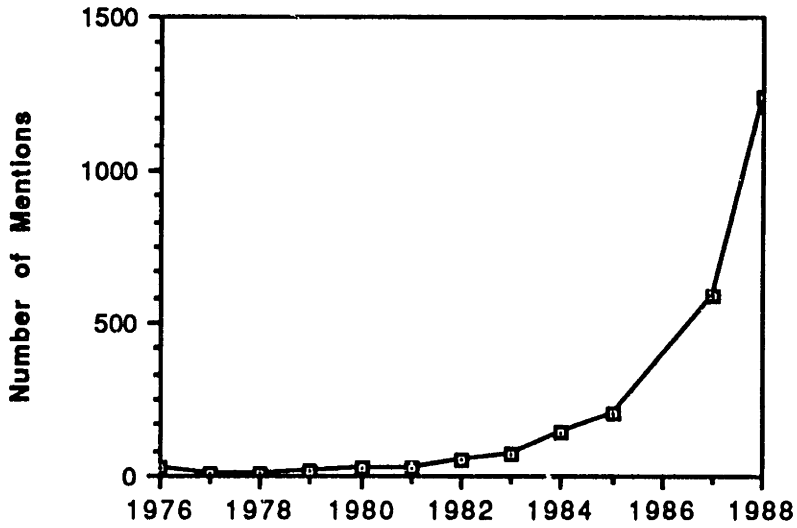


Figure 6.2
DAWN Reported Medical Examiner Mentions for Cocaine and Heroin in Combination



The second event is the global glut in opium production that has caused the retail price of heroin to decline significantly even as retail purities have soared.[9] This could induce more people to use heroin because it is cheap enough and pure enough to smoke,[10] and people who start smoking heroin may eventually inject it to achieve a more intense high,[11] particularly if renewed enforcement pressure drives the retail price and purity back to their previous levels.

The third event, the advent of AIDS, is the subject of this chapter. Although in some respects the three events are related, this chapter focuses exclusively on the third. In a sense the models presented here are of a hypothetical world that never existed, one in which IV drug use was in steady state before being disturbed by a single event, the advent of AIDS. Such models are useful, however, because they can help distinguish the effects of various factors. In principle, one could estimate the effect of the other factors by observing the actual population size and adjusting for the effect of AIDS predicted by the model.

Even between 1975 and 1985 there was some turnover in the IV drug using population. Some people began injecting. Others stopped, either because they quit injecting, died, or left the area of interest.

Little is known about what determines the aggregate rate at which people begin to inject drugs. Demographics and availability are clearly important, but so are such difficult to quantify factors as tastes, expectations about punishment, and the prevailing street wisdom about adverse health effects. It will be assumed here that the rate at which people join the population of IV drug users (the recruitment rate) is proportional to the number of current users raised to some power ν, where ν is nonnegative and strictly less than one.

The principal justification for this is that new users rarely begin injecting on their own; they are almost always introduced to the practice by a friend. Hence the number of people who are likely candidates to begin injecting increases with the number of current users.

This suggests that ν may be close to one. However, the majority of such initiations occur soon after the friend has begun

---

[9]National Narcotics Intelligence Consumers Committee, 1988; Isikoff, 1989a.

[10]*Time*, (1990) reports that "for the past several years, less diluted heroin from Southeast Asia that can be smoked has been widely available on the streets of New York, Boston, and other cities."

[11]Power (1988) and Newcombe (1989).

using IV drugs,[12] so the number of people who began injecting in recent months is probably as important as the total number of users.

Another view is that ν is much closer to zero because drugs in general, including heroin, are widely available to people who might begin to inject.[13] Hence anyone who is predisposed to inject drugs will, and the recruitment rate is really determined by demographics and preferences. Since the number of people at risk does not fluctuate widely and "all other things," including tastes, are assumed to be constant, this view suggests that ν is quite small, and therefore the number of recruits remains fairly constant, regardless of the size of the drug population.

This second argument is predicated on the assumption that as long as the number of current users stays within some range, drugs will continue to be widely available. Or, to put it another way, while there may be some critical mass of users necessary to support a distribution network that makes drugs widely available, that critical mass is small enough that one can safely assume it will be present for the foreseeable future. Fortunately, the analysis does not depend on knowing the precise value of ν.

In contrast, the rate at which users stop injecting is probably roughly proportional to the number of users. One argument against this is that there are "industry-wide economies of scale" in drug consumption.[14] Presumably as the number of users increases, so will the number of dealers, and finding a "connection" may become easier as the size of the drug market increases, thereby decreasing search time costs. Also, other things held constant, the more users there are, the less enforcement pressure there is per user. These "economies" suggest that the exit rate may rise less than proportionally with the number of current users. This is probably a second order effect, however, so it will not be incorporated into the model below. Including it would have the same effect on the steady state solution as an increase in ν.

This discussion suggests that a (greatly simplified) model of the population dynamics of the IV drug using population is

$$\frac{dN(t)}{dt} = c\,N(t)^{\nu} - \mu\,N(t) \qquad (6.1)$$

---

[12]Kaplan, 1983a.

[13]According to the High School Senior Survey about one-quarter of high school seniors report that it would be easy for them to obtain heroin. This has not changed much since 1976. U.S. Department of Health and Human Services, 1988c, pp.153-157.

where

c is the recruitment rate proportionality constant,
μ is the rate at which current users exit the population, and
N(t) is the size of the IV drug using population at time t.

Solving this equation[15] gives

$$N(t) = \left[\frac{c}{\mu} + \left(N(0)^{1-\nu} - \frac{c}{\mu}\right) e^{-\mu(1-\nu)t}\right]^{1/(1-\nu)} \qquad (6.2)$$

where N(0) is the population size at time 0. Over time the population approaches a steady state size of

$$N(\infty) = \left(\frac{c}{\mu}\right)^{\frac{1}{1-\nu}}. \qquad (6.3)$$

Not surprisingly the greater the rate at which new injectors are recruited, the larger the steady state population will be, and the greater the rate with which current users exit, the smaller the steady state population will be. Also, the closer ν is to 1, the more sensitive the steady state population will be to these parameters.

The author is not aware of any studies that have estimated the recruitment rate of IV drug users directly, but Kleiman and Mockler[16] did try to estimate the recruitment rate for heroin users and used that as an estimate of the recruitment rate for IV drug users.

They note that between 1980 and 1986 about 1% of high school seniors surveyed reported at least experimenting with heroin. Nationally this would be about 40,000 students. Adjusting for the fact that the survey excludes high school dropouts and that the median age of first heroin use is about 18, Kleiman and Mockler estimate that about 100,000 people per year begin using heroin.

Some heroin users do not inject the drug and some people inject drugs others than heroin, so recruitment of heroin users may be larger or smaller than the recruitment of injectors. However, most heroin users inject, heroin is the most frequently injected drug, and there are indications that at least among the population covered by the National Household Survey on Drug Abuse the number of

---

[14]This possibility will be discussed in Chapter 7.

[15]The author thanks Hongtao Zhang for showing him the solution.

[16]Kleiman and Mockler, 1988, pp.8-9.

people who report having ever used heroin is comparable to the number of people who report having ever injected drugs.[17] So it will be assumed that there are 100,000 recruits per year in steady state.

Not surprisingly no one knows for certain how many IV drug users there are, and several authors[18] have raised serious questions about some of the techniques currently used to estimate this number. However, numbers around 500,000 are commonly cited.[19] Official estimates are often considerably higher, perhaps because the former figure may refer just to heroin addicts. For instance, the U.S. Public Health Service[20] estimates that there may be as many as 1,500,000 IV drug users (divided into 750,000 "regular" and 750,000 "occasional" users). It will be assumed here that the steady state number of IV drug users before the advent of AIDS was 750,000.

These numbers imply the exit rate parameter was

$$\mu_B = \frac{100{,}000}{750{,}000} \cong 0.133 \tag{6.4}$$

and, for any given value of ν, the recruitment rate proportionality constant was

$$c_B = \frac{100{,}000}{750{,}000^{\nu}}. \tag{6.5}$$

The subscript B is used to denote "before" AIDS.

An exit rate of $\mu_B = 0.133$ is reasonable. If the length of drug careers is exponentially distributed, it suggests that the average career length is about 7-8 years. Although some people use heroin for longer periods, many use for less, so a 7-8 year average is at least plausible.

Consider now how the advent of AIDS would affect the number of IV drug users. Quite conceivably it would change the recruitment rate and exit rate parameters. This simple model will assume that

---

[17]The U.S. Department of Health and Human Services (1989a, pp.102-103) reports the numbers are 1.9 and 2.5 million respectively.

[18]See the discussion in Section 2.1.

[19]For example by Kaplan (1983b), Isikoff (1989a), Martz (1990), and Harwood et al. (1984, p.88). The most recent national household survey estimated that among the population surveyed 492,000 (95% confidence interval of 335,000-734,000) had injected drugs within the last year (U.S. Department of Health and Human Services, 1989c, p.103).

[20]U.S. Public Health Service, 1986.

these two parameters change instantaneously from μ_B and c_B to μ_A and c_A respectively (A for "after" or "AIDS").

Clearly this is a simplification; behavior takes time to change. However, the steady state population size is determined only by the ultimate values of the parameters (assuming they themselves stabilize once behavior has fully adjusted), and the principal result is a steady state result. Also, there is evidence that both HIV infection[21] and awareness of AIDS[22] spread very rapidly among IV drug users. Hence, although the changes in behavior represented by the changes in the parameter values would by no means occur instantaneously, they might occur over a time span that is relatively short compared to the time it would take the population size to reach the new steady state.

Assuming that the two parameters change values instantaneously, and taking t = 0 to be the point at which the parameters change, the population size as a function of time would be

$$N(t) = \left[\frac{c_A}{\mu_A} + \left(\frac{c_B}{\mu_B} - \frac{c_A}{\mu_A}\right) e^{-\mu_A(1-\nu)t}\right]^{1/(1-\nu)}, \qquad (6.6)$$

and the new steady state population size would be

$$N_A(\infty) = \left(\frac{c_A}{\mu_A}\right)^{\frac{1}{(1-\nu)}} = \left(\frac{c_A}{c_B}\frac{\mu_B}{\mu_A}\right)^{\frac{1}{(1-\nu)}} \left(\frac{c_B}{\mu_B}\right)^{\frac{1}{(1-\nu)}} = \left(\frac{c_A}{c_B}\frac{\mu_B}{\mu_A}\right)^{\frac{1}{(1-\nu)}} N_B(\infty). \qquad (6.7)$$

Not surprisingly, this model suggests that the relative sizes of the recruitment and exit rate parameters before and after the advent of AIDS determine how AIDS will affect the steady state number of IV drug users. The next few paragraphs discuss what the values of c_A and μ_A might be.

First consider how the advent of AIDS might affect recruitment. One would certainly think recruitment would be lower with AIDS ($c_A < c_B$) because the risk of contracting AIDS is a significant disincentive to injecting drugs.

Some might speculate that c_A could actually be greater than c_B if people respond to the threat of AIDS by making sterile needles

---

[21]Kaplan, 1989.

[22]Friedman, Des Jarlais, Sotheran, et al. (1987) and Ginzburg (1989, p.69) report that even before there were any official efforts to educate IV drug users, almost all of them knew of AIDS and 90% knew it can be transmitted by sharing needles.

widely available as has been done on a small scale by New York City and by private individuals in Boston and Portland, Oregon. This fear seems unjustified, however, because Amsterdam has had a large-scale syringe exchange program, and it apparently has not led to an increase in the number of IV drug users or to a decrease in the number of users seeking treatment.[23]

There is little evidence about the magnitude of any decrease in the recruitment rate proportionality constant. One bit of relevant data is that even as early as 1983, 16% of a group of treatment clients surveyed believed that "AIDS has influenced 'many' people who had never used drugs to stay away from drugs." Another 32% said "some" had been so influenced.[24]

However, the most recent high school surveys do not show any change in the fraction of high school seniors who have experimented with heroin.[25] This apparent lack of decline is particularly significant because the surveys show that the prevalence of drug use in general has been declining steadily and substantially. Reported marijuana use, for example, has dropped every year since 1979.

On the other hand, AIDS' full impact on the recruitment rate may not have been realized yet. New prevention programs motivated by concern about AIDS are still being implemented. Also, most people today are aware of AIDS, but it is likely that fewer knew personally someone who died of AIDS than will be the case in the future, and as Power[26] notes, that experience can induce behavioral changes when conventional education programs do not.

In summary, intuitively one would expect AIDS to reduce the recruitment rate proportionality constant, but at present there is not a great deal of evidence to support this conjecture. To be conservative, the assumption will be made that $c_A = c_B$, i.e., that AIDS does not appreciably affect recruitment.

It is almost certain, however, that AIDS will affect the exit rate. One, but by no means the only, reason for this is simply that infected users will die of AIDS.

Estimates of the mean and median of the incubation period for people who were infected by blood transfusions and homosexual

---

[23]Committee on AIDS Research (1989, p.202) and Buning et al. (1988).

[24]Ginzburg, 1989, p.70.

[25]U.S. Department of Health and Human Services (1989a) gives information through the 1988 survey. Isikoff (1990) indicates that most trends continued with the 1989 survey.

[26]Power, 1988, p.85.

contact are usually between 7.3 and 9.8 years.[27] There is no guarantee that the incubation period is the same for IV drug users, but the author is not aware of any similar studies for IV drug users. In general IV drug users are in worse health than the public at large, so it seems unlikely that they would live longer on average after being infected than transfusion recipients or homosexual men.

This suggests that the exit rate for an infected user is at least $\mu_B + 1/8$. The exit rate may be higher if infected users who know they are infected stop injecting to avoid infecting others or to avoid multiple exposure to the virus. Likewise, it may be higher if infected users who do not know they are infected stop injecting in the (mistaken) hope that they can avoid getting infected.

AIDS would also presumably increase the exit rate for uninfected users. Some uninfected users will stop injecting to avoid getting the virus, although this may not happen as often as one would expect. While there is convincing evidence that AIDS has induced risk reducing behavioral changes among IV drug users,[28] using sterile injection equipment and sharing with fewer people seem to be more common adaptations than injecting less often, although the last has been reported.[29] Injecting may also be reduced indirectly because as Des Jarlais and Hunt[30] point out, "both New York and New Jersey have increased their drug abuse treatment capabilities as a means of preventing AIDS," and other organizations will likely do the same if they have not already done so.

The exit rate μ_A is a weighted combination of the exit rates for infected and uninfected users, with the weights determined by the prevalence of infection. Of course the prevalence changes over time, but since the simplifying assumption has been made that the parameter values change instantaneously from their before AIDS values to their after AIDS values, the appropriate weights for the steady state exit rate would be determined by the steady state prevalence, π.

A lower bound for μ_A can be obtained by considering only increases in the exit rate directly attributable to infected users dying of AIDS. So

---

[27]Estimates in this range are given by Bacchetti and Moss (1989), Kalbfleisch and Lawless (1988), Lui, Darrow, and Rutherford (1988), and Medley et al. (1987).

[28]Committee on AIDS Research, 1989.

[29]Des Jarlais and Hunt (1988), Ginzburg (1989, pp.69-70), Friedman, Des Jarlais, Sotheran (1989, pp.204-205), and Des Jarlais, Friedman, Sotheran, and Stoneburner (1988, p.168).

[30]Des Jarlais and Hunt, 1988, p.4.

$$\mu_A \geq \mu_B + \pi/8. \qquad (6.8)$$

This seems unduly conservative though. Since there is evidence that the steady state prevalence will be quite high,[31] and in light of the arguments above, it will be assumed that

$$\mu_A = \mu_B + 1/8. \qquad (6.9)$$

In that case, even if AIDS has no effect on recruitment, the model predicts that AIDS will reduce the size of the IV drug using population from

$$N_B = \left(\frac{c_B}{\mu_B}\right)^{\frac{1}{1-\nu}} = 750{,}000$$

to

$$N_A = \left(\frac{c_A}{\mu_A}\right)^{\frac{1}{1-\nu}} \approx \left(\frac{c_B}{\mu_B + \frac{1}{8}}\right)^{\frac{1}{1-\nu}} = \left(\frac{\mu_B}{\mu_B + \frac{1}{8}}\right)^{\frac{1}{1-\nu}} \left(\frac{c_B}{\mu_B}\right)^{\frac{1}{1-\nu}}$$

which is approximately

$$\cong \left(\frac{1}{2}\right)^{\frac{1}{1-\nu}} \left(\frac{c_B}{\mu_B}\right)^{\frac{1}{1-\nu}} \qquad (6.10)$$

because $\mu_B \approx 0.133$.

This result is startling because it suggests that no matter what ν is, AIDS will reduce the size of the IV drug using population by at least 50%. Before the advent of AIDS people left the population at a rate of about $1/8$ yr^{-1}. A reasonable guess is that AIDS has created a new set of reasons for exiting the population at a rate of about $1/8$ yr^{-1}, so roughly speaking AIDS has approximately doubled the exit rate. If recruitment remained constant, this would halve the steady state population size. But if recruitment increases with the population size (i.e. if $\nu > 0$), this decrease will reduce recruitment, which will further reduce the steady state population size.

A key question is how long it will take for the population to reach the new steady state. The time constant for $N(t)^{1/(1-\nu)}$ is the reciprocal of $\mu(1-\nu)$, and one can show that $N(t)$ approaches its new equilibrium value roughly as quickly as $N(t)^{1/(1-\nu)}$ does. Hence if the

---

[31] Kaplan, 1989.

behavior represented by the recruitment and exit rate parameters changed instantaneously, one would expect to observe the bulk of the change in the population size within one to two time constants, i.e. within roughly 5 to 15 years as long as ν is not more than $1/2$.

Of course behavioral changes do not occur instantaneously, and although HIV prevalences are very high in some areas, some cities are still in the early stages of the epidemic. Hence it might be considerably longer before a new steady state is reached.

One could be excused for being skeptical about such a dramatic prediction. After all it is based on a simple descriptive model. To help alleviate such concerns, the next section develops a more detailed model that explicitly considers needle sharing, the principal means by which the HIV virus is transmitted among IV drug users. Somewhat surprisingly it suggests that Equation 6.3 gives a fair estimate of the steady state population, and hence that a 50% reduction in the steady state population size may be a reasonable prediction.

6.2 A Model of Needle Sharing

The result above is not completely persuasive because a model that describes how AIDS affects the size of the IV drug using population should explicitly acknowledge needle sharing behavior. Kaplan introduces such a model.[32] It focuses on "shooting galleries" which are believed to play a key role in the spread of AIDS.[33]

Shooting galleries are places where users rent injection equipment. Conventional shooting galleries are most common in big cities in the Northeast, but other forms of needle sharing have similar effects on the transmission of the HIV virus.[34]

Although needle sharing is very common,[35] it is not the only way AIDS spreads among IV drug users. IV drug use and "risky" sexual practices are often both integral parts of a general lifestyle,[36] but sexual transmission is not addressed in this model.

---

[32]Kaplan, 1989.

[33]Marmor et al. (1987) and Des Jarlais, Friedman, and Stoneburner (1988).

[34]As Des Jarlais, Friedman, Sotheran, and Stoneburner (1988) suggest for "houseworks", Feldman and Biernacki (1983) describe for residential hotels in San Francisco, and Mata and Jorquez (1988) describe for needle sharing among Mexican-Americans.

[35]Friedland et al., 1989.

[36]Committee on AIDS Research (1989), Feldman and Biernacki (1988), and U.S. Department of Health and Human Services (1987).

In its basic form Kaplan's model includes quantities such as rates of sharing injection equipment, the ratio of addicts to injection equipment in the population, the infectivity of HIV transmitted by shared injection equipment, the likelihood that infectious equipment is "flushed" by the blood of an uninfected user, and the duration of needle-sharing activity by HIV-infected addicts.

However, the model takes the number of IV drug users as a constant. It assumes that whenever a user exits the population another is recruited. This assumption, although clearly imperfect, is common in epidemiological modeling.[37] If the disease spreads rapidly relative to its incubation time, as is the case with AIDS, it is not a gross oversimplification when the quantity of interest is the fraction of the population infected at any given time. Clearly though, one cannot assume the population size is fixed if that is the variable of interest, so this section develops an open population version of the Kaplan model.

Let N(t) represent the number of IV drug users at time t. At any moment some number I(t) of these users are infected with HIV and the remaining U(t) = N(t) − I(t) are not infected. The rates of growth and decline of I(t) and U(t) depend on a third quantity, $\beta(t)$, the fraction of needles that are infectious at time t.

Making some simplifying assumptions allows one to relate these quantities through a simple model. The assumptions are discussed in detail by Kaplan[38] and are more briefly reviewed here.

(1) All needle sharing occurs in one of m identical and indistinguishable shooting galleries. Shooting galleries are defined as places where users sequentially use one set of injection equipment. Physical shooting galleries that have more than one set of equipment are modelled as several different shooting galleries.
(2) Each user visits shooting galleries according to a Poisson Process with rate λ, independent of the actions of the other users.
(3) Injection equipment becomes infectious when it is used by an infected user, and it remains infectious until it is "flushed" by an uninfected user.
(4) If an uninfected user injects with equipment that is not infectious the user remains uninfected and the equipment not infectious. If, on the other hand, an uninfected user injects with infectious equipment, he or she becomes infected with probability α, and with probability θ the equipment ceases to be infectious (the user "flushes" the

[37] Anderson and May, 1982.
[38] Kaplan, 1989.

equipment). The events that the uninfected user becomes infected and the equipment ceases to be infectious are assumed to be independent.
(4) Users can only become infected by sharing injection equipment.
(5) Variability in $\pi(t)$, the fraction of infected addicts addicted at time t, can be ignored.
(6) As before, people join the population of IV drug users at a rate which is proportional to the number of current users raised to the power ν with $0 \leq \nu < 1$. New users are not infected.
(7) Uninfected users exit the population at rate μ_1. Assuming that $I(t) > 0$, μ_1 would be larger than the exit rate before the advent of AIDS because of the deterrence effects described above. Infected users exit the population at rate $\mu_2 > \mu_1$.

Assumptions 1 and 2 imply "perfect mixing," i.e. that all users visit shooting galleries at the same rate, and they are as likely to visit one gallery as another. This is unlikely to be the case in practice, and Kaplan extends his basic needle sharing model to include variability in the rate with which users visit shooting galleries. He finds that the greater the heterogeneity, the more rapidly the prevalence approaches steady state but the lower that steady state prevalence is. Nevertheless heterogeneity is not considered here because it complicates the analysis and so little is known about the distribution of rates with which users visit shooting galleries that there is little hope of taking advantage of such a refinement.

Kaplan also extends his basic model to include sterilizing needles before use and the inactivation of the virus in a needle over time. These extensions are not addressed here because in this context they add little except algebraic complexity.

Table 6.1 summarizes the notation used in this model.

The first step in formulating the model is to describe how the probability that a randomly selected needle is infectious ($\beta(t)$) varies over time. During any sufficiently short time interval Δt one of three things can happen to a needle. It can be used by someone who is infected (Event A); it can be used by someone who is not infected (Event B); or it may not be used at all (Event C). By the Poisson assumption, one can neglect events involving more than one use of the needle.

Table 6.1:
Notation Used In Needle Sharing Model

m = the number of shooting galleries
λ = the rate at which users visit shooting galleries
α = the probability an uninfected user who injects with an infectious needle will become infected
θ = the probability an infectious needle ceases to be infected after it is used by an uninfected user
c = the recruitment rate proportionality constant
$\mu 1$ = the exit rate for uninfected users
$\mu 2$ = the exit rate for infected users
ν = the exponent of number of users in the recruitment rate
$\beta(t)$ = the fraction of needles that are infectious at time t
$I(t)$ = the number of infected users at time t
$U(t)$ = the number of uninfected users at time t
$N(t)$ = the total number of users at time t

The $I(t)$ infected users choose one of the m identical needles with rate λ, so in a time interval of length Δt the probability that any given needle is used by an infected user is $P\{A\} = \frac{\lambda I(t)}{m}\Delta t$. Likewise the probability it is used by an uninfected user is $P\{B\} = \frac{\lambda U(t)}{m}\Delta t$. Hence the probability that no one uses the needle is just $P\{C\} = 1 - \frac{\lambda (I(t)+U(t))}{m}\Delta t$.

If the injection equipment was used between t and $t + \Delta t$ by someone who was infected, then with probability 1 the injection equipment is infectious at time $t + \Delta t$. The corresponding probability if no one used the needle is just $\beta(t)$, the probability the needle was already infectious at time t.

If someone who was not infected used the needle it will be infectious at time $t + \Delta t$ if and only if it was infected at time t and was not flushed by the user. This occurs with probability $\beta(t)(1 - \theta)$. Hence

$$\beta(t + \Delta t) = \beta(t + \Delta t|A)\ P(A) + \beta(t + \Delta t|B)\ P(B) + \beta(t + \Delta t|C)\ P(C)$$

$$= 1\left(\frac{\lambda I(t)}{m}\Delta t\right) + \beta(t)(1 - \theta)\left(\frac{\lambda U(t)}{m}\Delta t\right) + \beta(t)\left(1 - \frac{\lambda (I(t)+U(t))}{m}\Delta t\right).$$

So

$$\beta(t + \Delta t) - \beta(t) = \frac{\lambda}{m}[(1 - \beta(t))\, I(t) - \theta\, \beta(t) U(t)]\, \Delta t.$$

Dividing by Δt and taking the limit as Δt goes to zero gives

$$\frac{d\beta(t)}{dt} = \frac{\lambda}{m}[(1 - \beta(t))\, I(t) - \theta\, \beta(t) U(t)] \qquad (6.11)$$

Differential equations can also be written for I(t) and U(t). Uninfected users inject with rate λ. An injection exposes a previously uninfected user to infection if the needle was infectious (which happens with probability $\beta(t)$). The user becomes infected when exposed to an infectious needle with probability α, so the rate at which uninfected users become infected is $\lambda\, \alpha\, \beta(t)\, U(t)$. Together with Assumptions 6 and 7 this implies that

$$\frac{dI(t)}{dt} = \lambda\, \alpha\, \beta(t) U(t) - \mu_2\, I(t) \qquad (6.12)$$

and

$$\frac{dU(t)}{dt} = c\, (U(t) + I(t))^{\nu} - \lambda\, \alpha\, \beta(t) U(t) - \mu_1\, U(t) \qquad (6.13)$$

where c is the recruitment rate proportionality constant.

Equations 6.11-6.13 are a system of nonlinear ordinary differential equations that describe how the HIV virus spreads among IV drug users and how it affects the size of that population.

In principle one could substitute values for the parameters in Equations 6.11-6.13, choose appropriate initial conditions, and predict the course of the epidemic. Unfortunately there is considerable uncertainty about some of the parameter values. However, one can derive some conclusions without knowing those values exactly by concentrating on the steady state.

Let β, U, and I be the steady state values for $\beta(t)$, U(t), and I(t) respectively. In steady state the right hand sides of Equations 6.11-6.13 equal zero. The resulting system of three equations in three unknowns has two solutions. In the first no one is infected ($I(t) = \beta(t) = 0$) and the number of IV drug users is

$$U = N = \left(\frac{c}{\mu_1}\right)^{\frac{1}{1-v}}. \tag{6.14}$$

This is the situation before the advent of AIDS (or after the epidemic has died out).

The second solution, which is only valid if $\lambda\alpha > \mu_2\theta$, has

$$\beta = 1 - \frac{\mu_2\theta}{\lambda\alpha}, \tag{6.15a}$$

$$I = (\lambda\alpha - \theta\mu_2)\left[\frac{c(\mu_2 + \lambda\alpha - \theta\mu_2)^v}{\mu_2(\mu_1 + \lambda\alpha - \theta\mu_2)}\right]^{\frac{1}{1-v}}, \tag{6.15b}$$

$$U = \mu_2\left[\frac{c(\mu_2 + \lambda\alpha - \theta\mu_2)^v}{\mu_2(\mu_1 + \lambda\alpha - \theta\mu_2)}\right]^{\frac{1}{1-v}}, \tag{6.15c}$$

$$N = I + U = \left[\frac{c(\mu_2 + \lambda\alpha - \theta\mu_2)}{\mu_2(\mu_1 + \lambda\alpha - \theta\mu_2)}\right]^{\frac{1}{1-v}}, \quad \text{and} \tag{6.15d}$$

$$\pi = \frac{I}{N} = \frac{\lambda\alpha - \theta\mu_2}{\lambda\alpha + (1-\theta)\mu_2}. \tag{6.15e}$$

Equations 6.15a and 6.15e are the same as those given by Kaplan. That is, the predicted HIV prevalence and fraction of needles that are infectious are the same in the open and closed population models.

Note that

$$\left(\frac{c}{\mu_2}\right)^{\frac{1}{1-v}} < N = \left[\frac{c(\mu_2 + \lambda\alpha - \theta\mu_2)}{\mu_2(\mu_1 + \lambda\alpha - \theta\mu_2)}\right]^{\frac{1}{1-v}} < \left(\frac{c}{\mu_1}\right)^{\frac{1}{1-v}}. \tag{6.16}$$

That is, this more realistic model also predicts that AIDS will reduce the size of the steady state population, but not by as much as the first model predicted.

If μ_1 and μ_2 are much larger than $\lambda\alpha - \mu_2\theta$, then the second model predicts that AIDS will not dramatically affect the size of the steady state population. But if $\lambda\alpha - \mu_2\theta$ is much larger than μ_2 (and

hence is much larger than μ_1) this model predicts decreases almost as substantial as those predicted by the first model.

Kaplan[39] suggests values of λ = 1/week, α = 0.075, and θ = 0.25. With these values and μ_2 = 0.125 yr.$^{-1}$, $\lambda\alpha - \mu_2\theta \cong 3.84$ yr.$^{-1}$ which is considerably greater than μ_2. So if these values are reasonable then

$$N = \left[\frac{c(\mu_2 + \lambda\alpha - \theta\mu_2)}{\mu_2(\mu_1 + \lambda\alpha - \theta\mu_2)}\right]^{\frac{1}{1-\nu}} \cong \left(\frac{c}{\mu_2}\right)^{\frac{1}{1-\nu}}. \qquad (6.17)$$

Hence the second model supports the first model's prediction that if all other factors such as drug prices, enforcement policies, and so on do not change, AIDS will substantially reduce the number of IV drug users.

These models also give an expression for the reduction in recruitment. Specifically, recruitment will be reduced to

$$\left(\frac{1}{2}\right)^{\frac{\nu}{1-\nu}} \qquad (6.18)$$

of its former level. This reduction is more sensitive to changes in the parameter ν than the reduction in the steady state population size.

6.3 Summary

This chapter predicts that, barring changes in other relevant factors, AIDS will substantially reduce the size of the IV drug using population. It presents two models, one quite simple, the other somewhat more complex, that describe how AIDS spreads among this population. Both produce essentially the same algebraic expression for the change in population size. For reasonable parameter estimates the change is quite large, perhaps on the order of 50% or more, but uncertainty about the parameter values and simplifications inherent in the modelling process make more precise predictions difficult.

---

[39]Kaplan, 1989.

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.

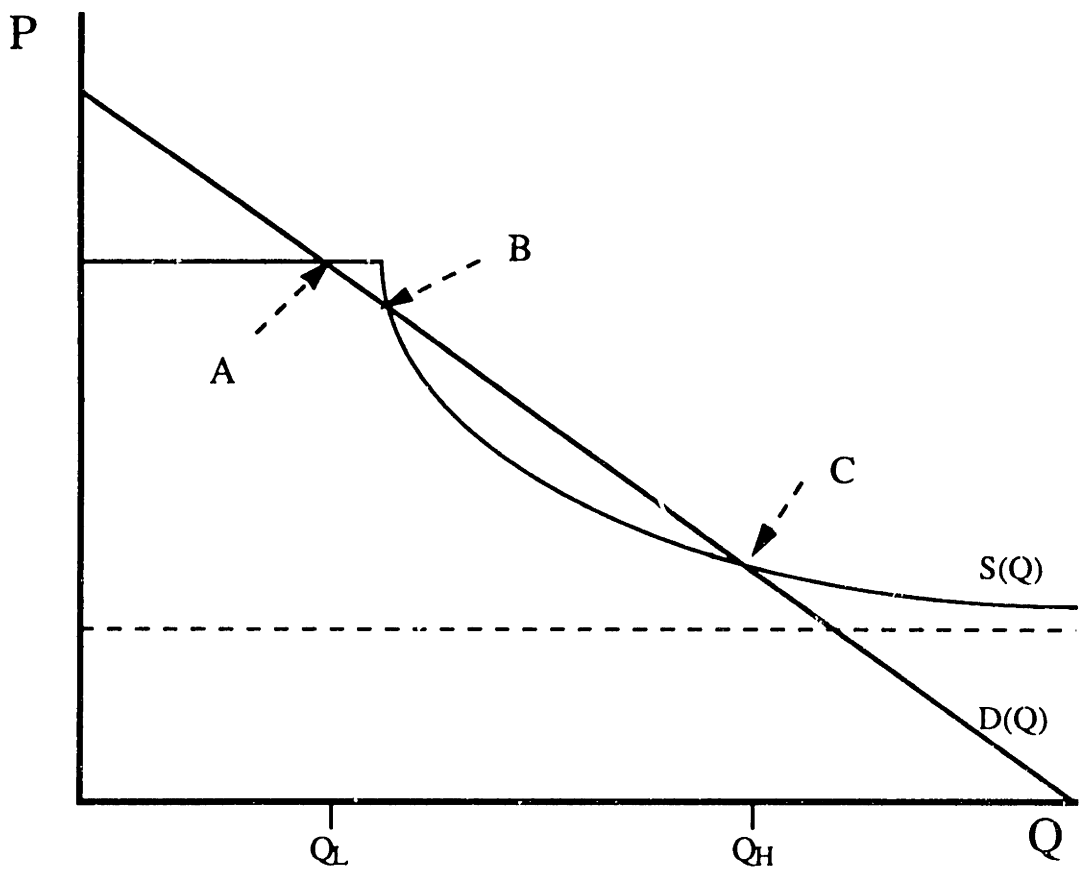
These non-price costs to users affect the market equilibria. The proper way to analyze them would be to draw the industry supply curve and then another curve above it representing the dollar

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.

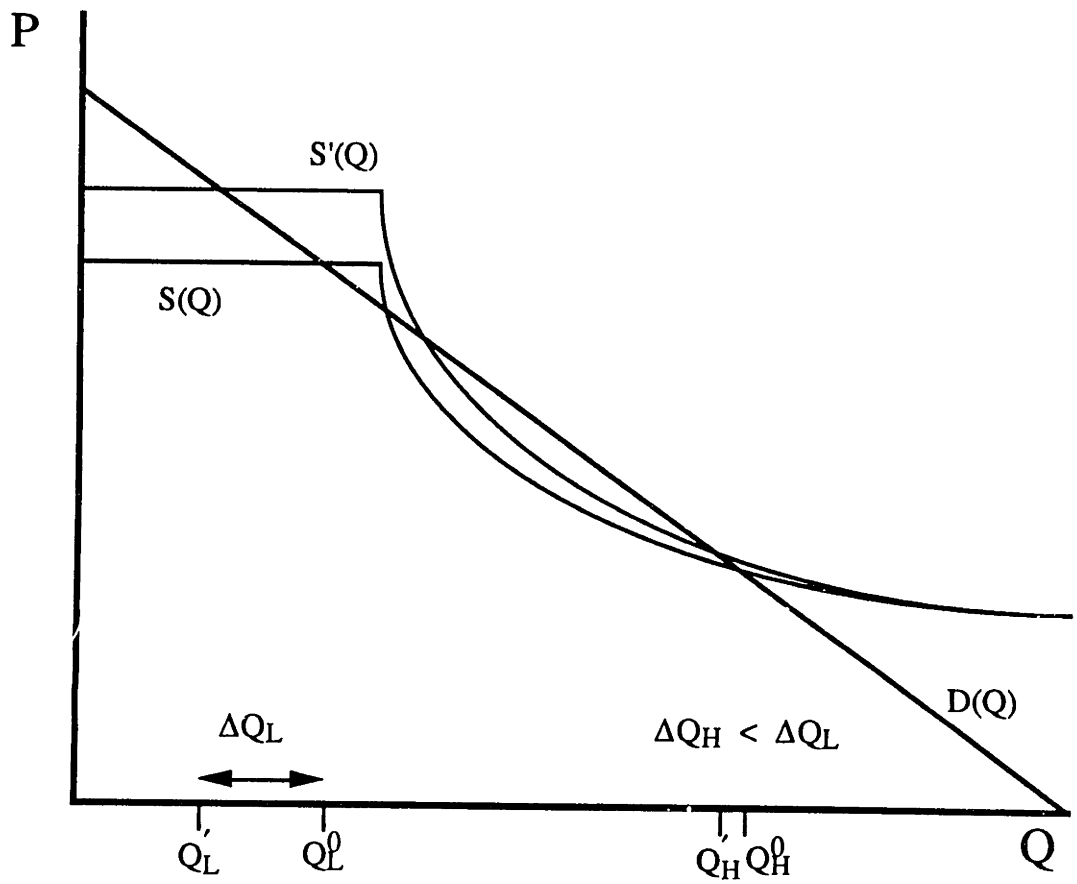
In either case there is an important lesson for communities that are still at point A. In the long run such communities would save considerably if they responded to increases in demand quickly, before the situation deteriorated from point A to point C.

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.

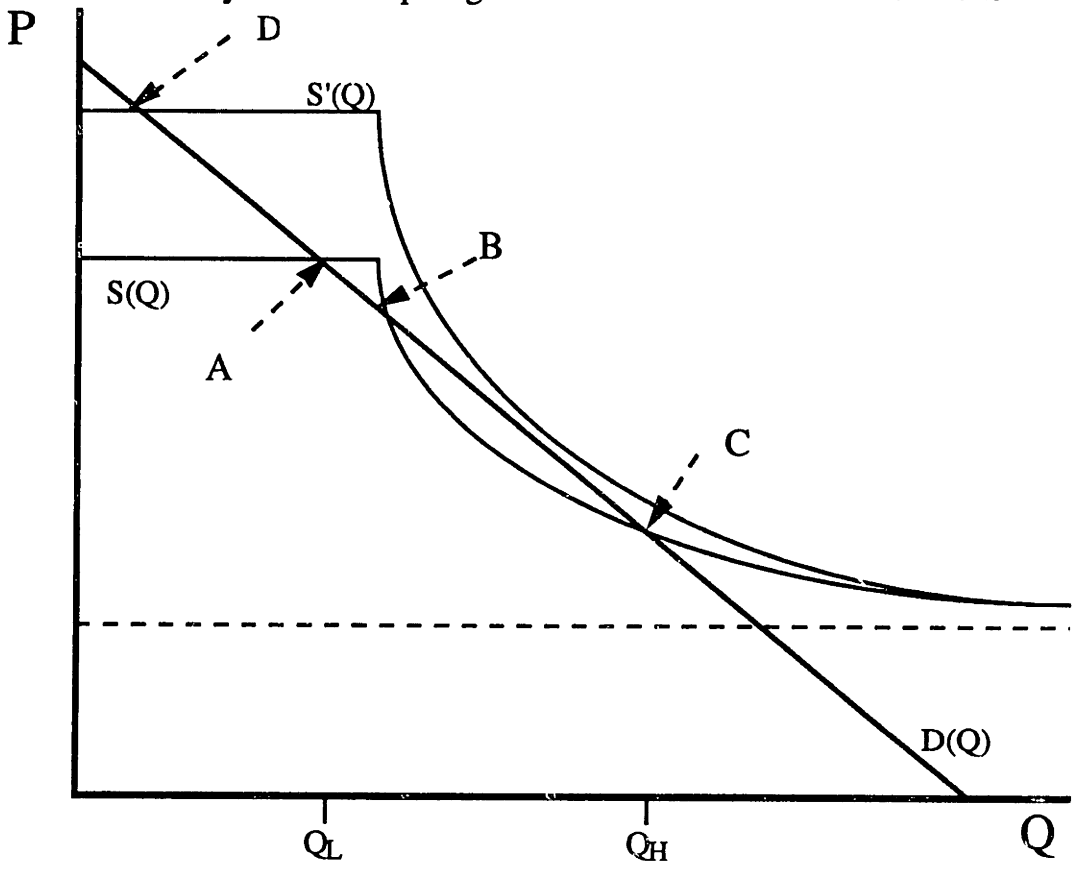
Hence, this model suggests that even if increasing the enforcement effort would be successful at point A, it may be relatively ineffectual if consumption has stabilized at the high levels described by point C.

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



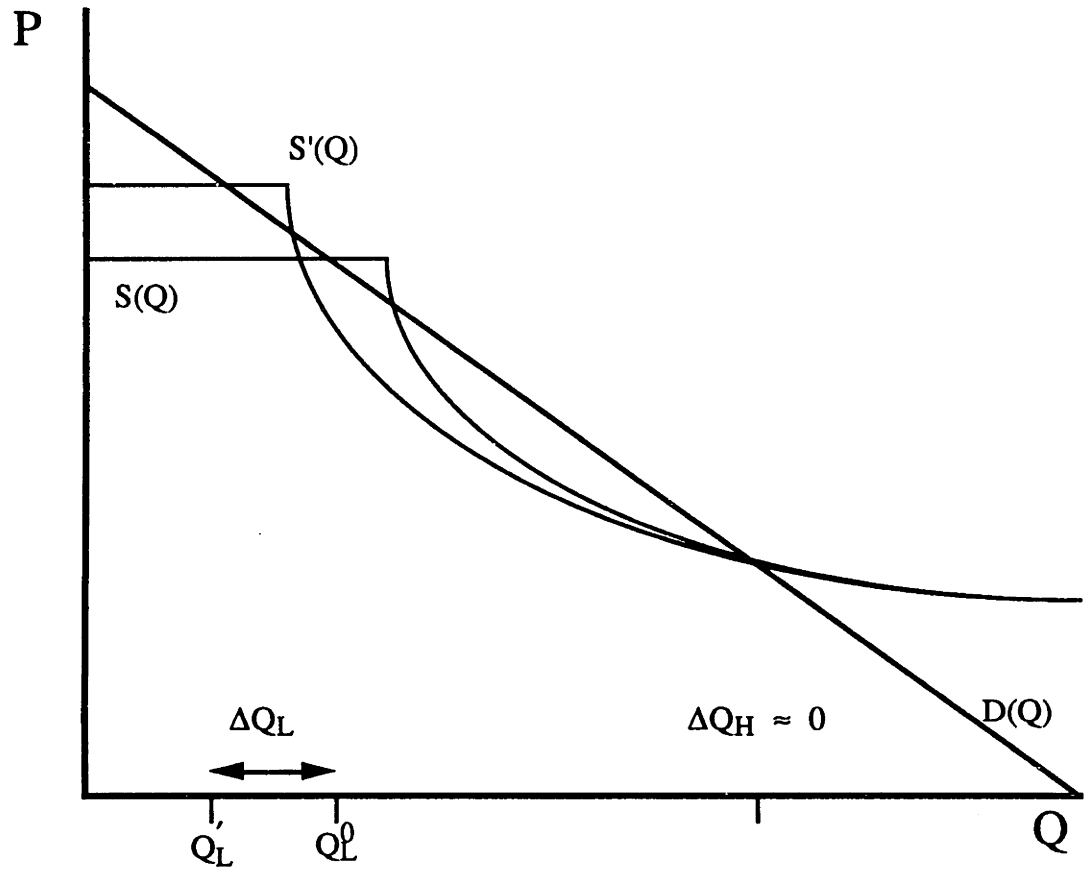
It is not clear how massive such an effort would have to be, nor how long it would have to last. The model is far too simplistic to even begin to shed light on such vital questions. Nevertheless it offers a glimmer of hope for enforcement. It is doubtful that such a massive crackdown could be achieved 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

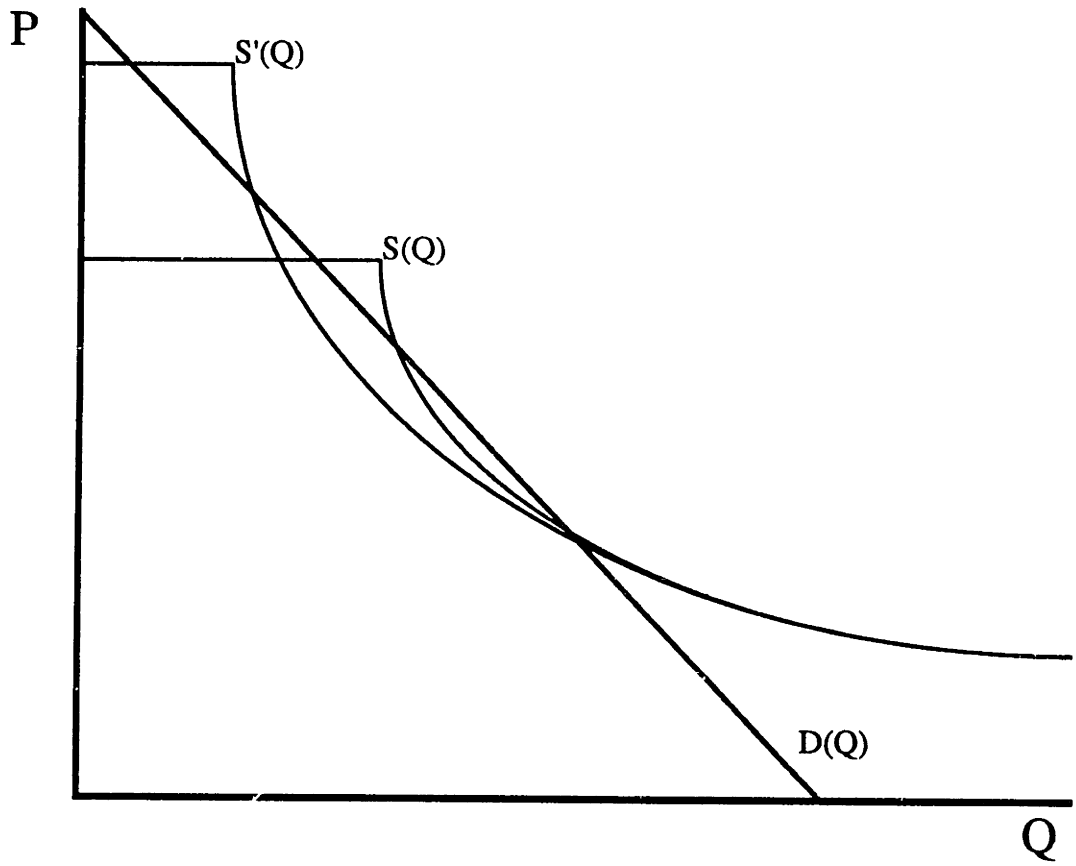


The other effect of minimum sentences would be to move to the left the kink in the supply curve representing the point at which the criminal justice system's punishment capacity becomes fully utilized. If the demand curve is sufficiently steep, the market were originally at point A, and point A were 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.

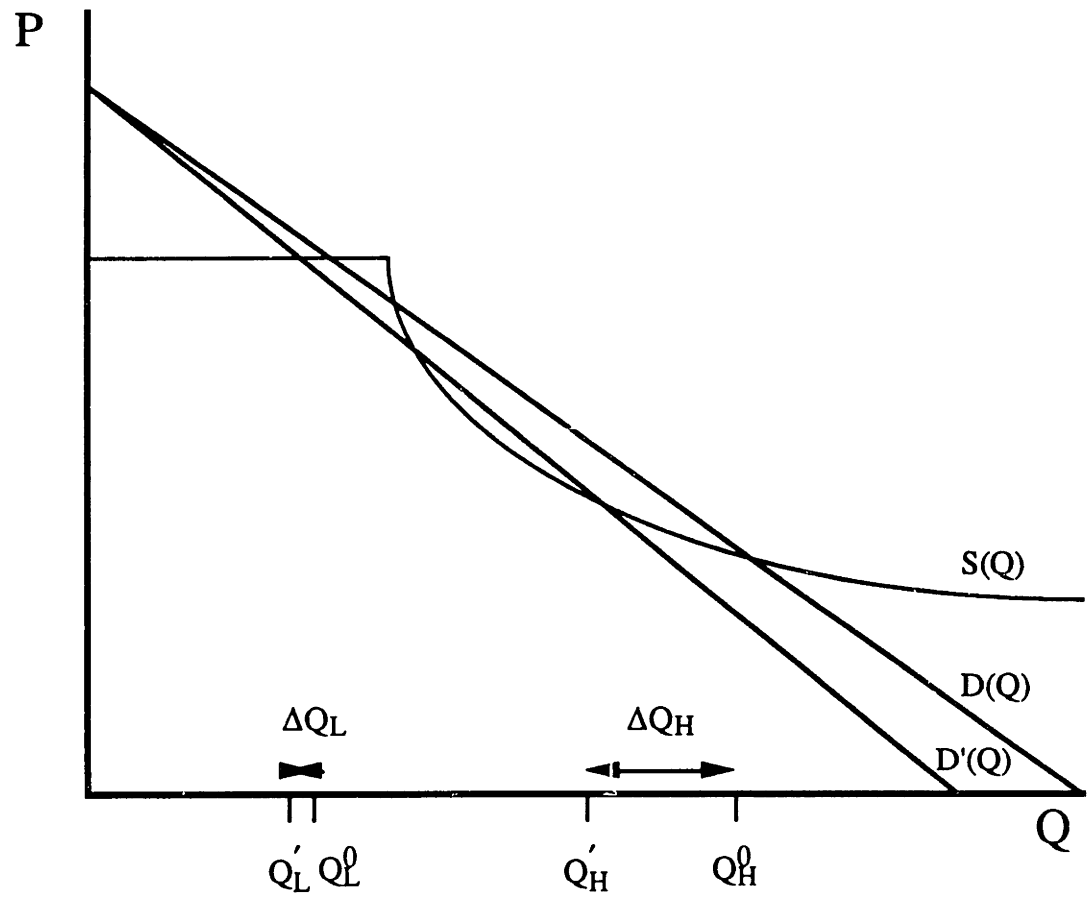
7.2.5 The Effectiveness of Demand Reduction

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

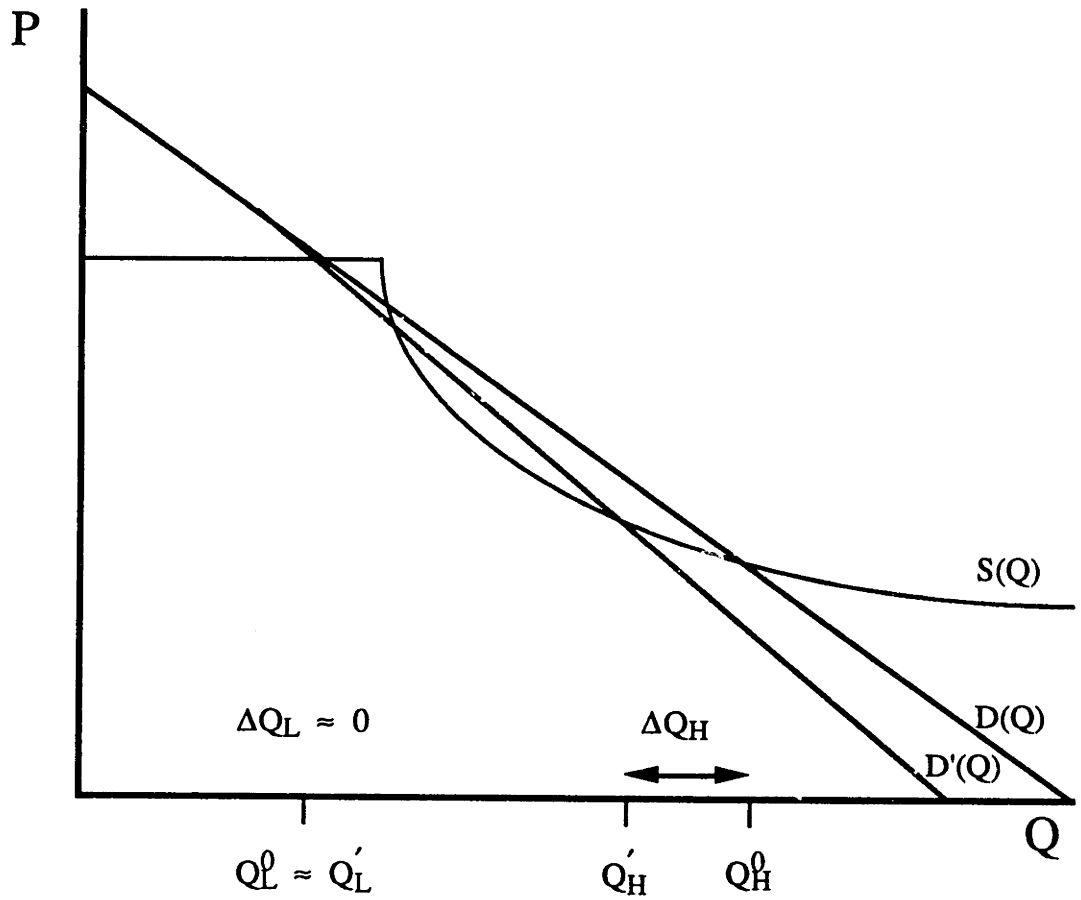


Education about the health effects of illicit drugs may be more effective with "casual" users and potential "casual" users that with truly committed users. The committed users are the ones who contribute demand even at high prices. "Casual" users, as the term is used here, are those who consume if and 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 societal 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

$$\begin{aligned} &R(x) = 1 + \alpha \\ &R(x^2) = (1 + \alpha)^2, \text{ and abstracting somewhat} \\ &R(x^\beta) = (1 + \alpha)^\beta. \end{aligned} \tag{7.1}$$

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

$$\begin{array}{ll} \text{Max} & [1 - (1 - p)^b]\, U(-C_F) + (1 - p)^b\, U(R(b)) \\ b \geq 0 & \end{array} \qquad (7.2)$$

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

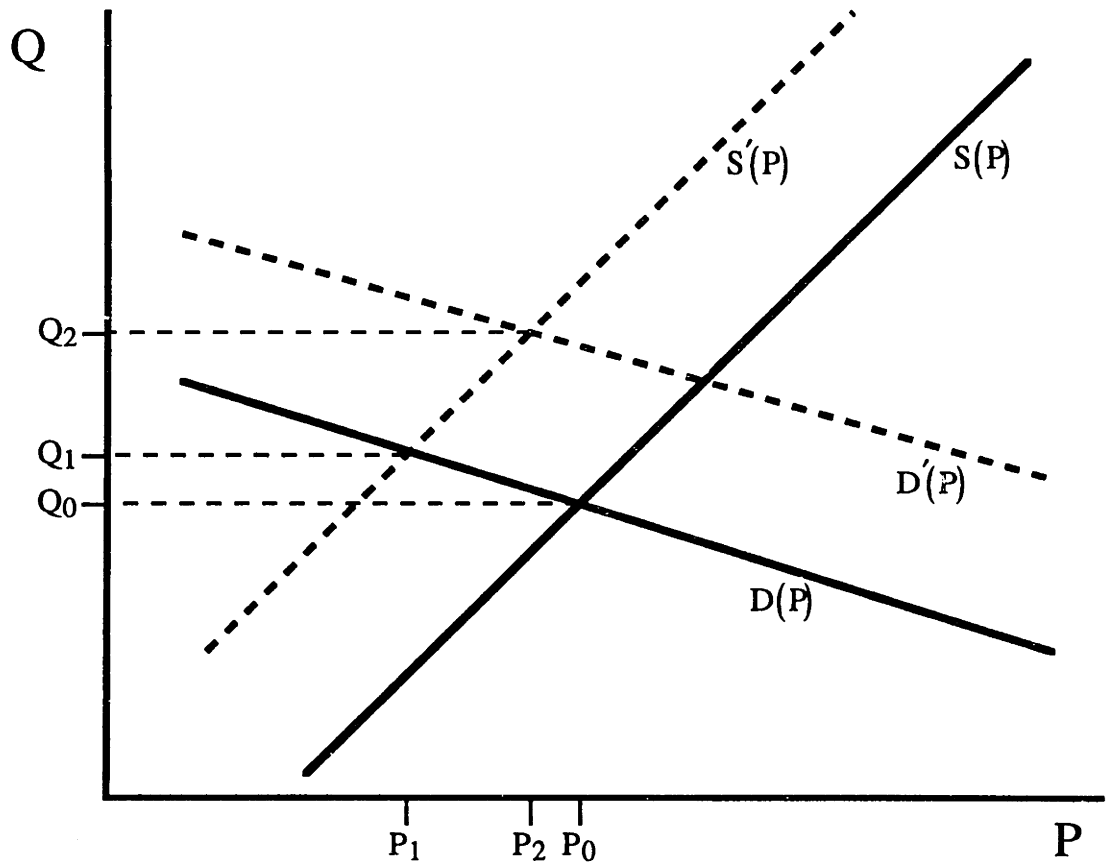
---

[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

$$\varepsilon = \frac{dQ(P)}{dP}\frac{P}{Q} = \beta_c + \frac{f_a}{1 - f_a}\beta_a < \beta_c \qquad (7.11)$$

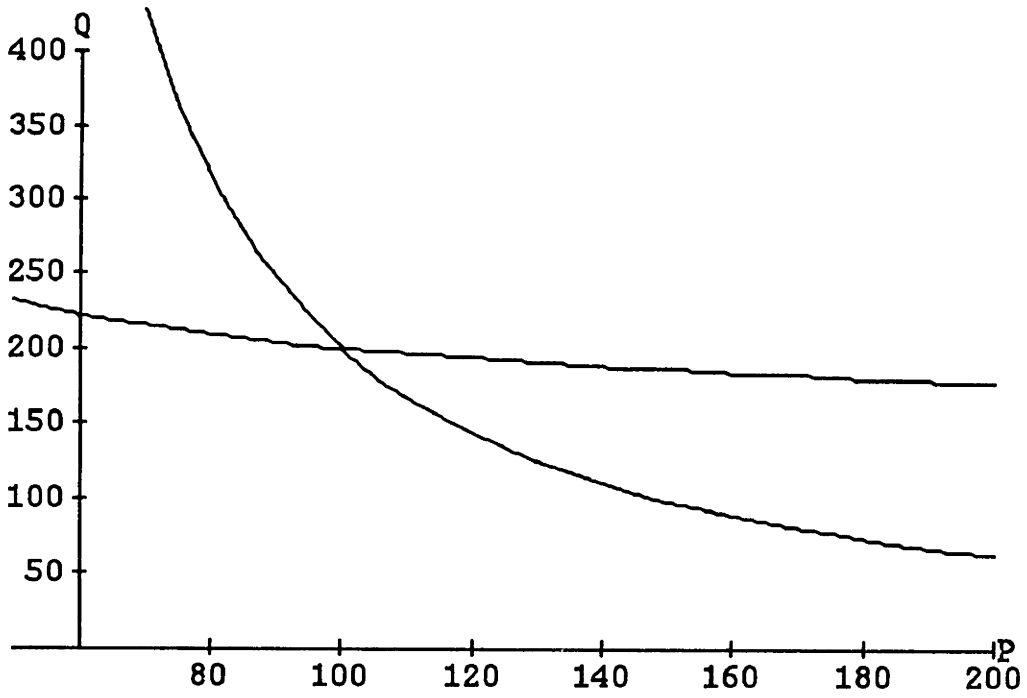
where $f_a = \alpha_a P^{\beta a}$ is the fraction of consumption accounted for by addicts.

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.

Chapter 8: Conclusions

This final chapter has two sections. The first reviews the principal results of the preceding chapters and makes recommendations for further work. The second tries to meld results of the individual chapters into three broad observations about "the drug problem": (1) The drug system is not linear and may, in some respects at least, be quite sensitive to exogenous changes; (2) Enforcement cannot "solve" the drug problem; and (3) Enforcement may be able to help "manage" the drug problem in ways that reduce the harm done by drug distribution, consumption, and enforcement.

8.1 Review and Recommendations for Further Work

8.1.1 Chapter 1

The first chapter argued that drugs cause and/or are associated with serious societal problems, that the drug system merits serious study, and that an Operations Research analysis can reasonably be expected to contribute to that study. It is hoped that this thesis as a whole confirms this expectation. If so, then an important conclusion from this work is that mathematical modelling can improve the collective understanding of the drug problem.

8.1.2 Chapter 2

Chapter 2 attempted to do two things. First it tried to alert the reader to the generally poor quality of the data and in particular to the existence of mythical numbers. Second, it briefly described sources of data that are available.

More is almost always better when it comes to a literature search. So developing a more comprehensive and more detailed catalog of relevant sources of information would be useful. It is with regard to the data themselves, however, that there is the greatest need for further work.

Collecting data on drug markets is inherently difficult, but there may well be room for improvement. Simple suggestions include adding new questions to existing surveys and surveying new subpopulations. More ambitious possibilities might include collecting and consolidating enforcement agencies' records, court documents, and testimony, or perhaps even observing some markets directly.

Understanding better how sampled populations represent or misrepresent larger populations is also vital. Such information could help one relate different data sets. Currently data is collected from high school seniors, college students, arrestees, treatment clients, and other special populations, but little can be, or at least seems to have been, done in the way of using two or more of these sources at once.

Finally it would clearly be valuable to debunk the mythical numbers. Scientific understanding advances by building on what is already known; if what is already "known" is wrong, then that building will be done on a foundation of sand. Sometimes it may be better to have 10 trustworthy numbers than to have 90 trustworthy numbers, 10 misleading numbers, and to not know which are which.

8.1.3 Chapter 3

The third chapter examined how changes in illicit drugs' prices at one level of the distribution network might affect their prices at subsequent levels. Conventional wisdom holds that increasing the import price by a dollar per unit will increase the retail price by roughly a dollar per unit. This view (called the "additive" model) implies that even if interdiction efforts were able to double the import price, it is unlikely that that would appreciably affect consumption because import prices are only a small fraction of retail prices.

Chapter 3 suggests an alternative, called the "multiplicative" model, according to which changes in import prices are passed along on a percentage basis. Doubling the import price would double the price at all subsequent levels and hence would probably have an appreciable effect on consumption.

Chapter 3 suggests the additive model is most likely to hold when costs depend primarily on the quantity of drugs transacted, as would likely be true when the drugs' value per unit weight or volume was not extremely high. On the other hand, when the drugs are very expensive, the distribution costs might depend more on their value than their weight, and one would expect the multiplicative model to hold. Hence one would expect the multiplicative model to be most nearly true for heroin and cocaine transactions in this country and the additive model to hold for marijuana both here and abroad and for cocaine and heroin in the source countries.

The limited price data that are available suggest that retail cocaine prices are considerably more responsive than the additive model predicts but somewhat less responsive than the multiplicative model predicts.

Perhaps the most valuable next step would be to obtain wholesale and retail price data for heroin and see how they are related. Another would be to compare changes in retail and wholesale enforcement to the corresponding changes in prices. Even a very predictable relationship between changes in wholesale and retail prices does not in any way establish causality. Another explanation would be that the factors determining the prices at both the wholesale and retail level changed in the same proportion. Testing this alternate hypothesis would be valuable.

8.1.4 Chapter 4

Chapter 4 developed a model, the balloon model, of how a so-called "open air" drug market would respond to local-level enforcement pressure. It assumes people will deal in a market if and only if the expected benefit of doing so is at least as great as the benefits available from other activities. An equilibrium exists when none of the dealers in the market have better alternatives elsewhere and none of the people who are not dealing are less well off than they would be if they were dealing. Since enforcement pressure can affect the desirability of dealing in a particular market, there is a relationship between the level of enforcement pressure and the equilibrium size of the market. Analyzing this relationship is the focus of the chapter.

Section 4.14 describes the model's results, so only the most important are reviewed here. The first of these is that the balloon model explicitly describes a positive feedback phenomenon. It suggests that the benefits of increasing enforcement may be nonlinear in the sense that the marginal benefit of another unit of enforcement (where benefit is measured in terms of reduction in the size of the market) is an increasing function of the amount of enforcement pressure, up to the point at which the market collapses. This gives an explicit, theoretical justification for a focused "crackdown" strategy. Equity considerations might favor distributing enforcement uniformly, but the balloon model strongly suggests that that would be less effective than concentrating on one market at a time.

The balloon model also supports the use of the balloon metaphor. Pressing down gently on a balloon will deform it, but the balloon springs back as soon as the pressure is removed. If one presses hard enough, however, the balloon will burst, and it will not spontaneously inflate again when the pressure is removed.

The same may be true of drug markets. Applying a modest amount of enforcement pressure may temporarily depress sales, but

it probably will not have any lasting effect once the pressure is removed. On the other hand, if sufficient pressure were applied to collapse a particular market, then most of that pressure could be subsequently removed without having the dealing return.

Extending the balloon model to include the interaction between different markets within the same city would be valuable. The key question is displacement. When enforcement pressure shrinks one market, to what extent will dealers and customers move to other pre-existing markets and/or form new markets?

Another important extension would be to model the dynamic response of the market. Currently the balloon model looks at the relationship between enforcement pressure and market size in steady state, but questions about the timing and duration of crackdowns are vital.

A third direction for further work would be to model more explicitly the effects of enforcement. Currently enforcement pressure is treated as a single exogenous parameter that directly affects dealers' utility. Clearly enforcement authorities' total impact on the dealers' utility depends on other factors, including the number of dealers, as well as the level of resources committed. Also enforcement can be directed at customers as well as dealers. More work could be done on modeling enforcement against customers and evaluating what combination of pressure against dealers and against customers is best.

8.1.5 Chapter 5

Chapter 5 examines how punishment policies affect the behavior of drug market participants, particularly so-called controlled users. For the purposes of the chapter a punishment policy is defined as the punishment one can expect to receive as a function of the quantity possessed at the time of arrest.

Somewhat surprisingly the model suggests that it is not always the case that giving the harshest possible sanction irrespective of the quantity possessed yields the lowest level of consumption. Under such a scheme people who still wish to consume have an incentive to buy very large quantities infrequently. In doing so they avoid "paying" the nonmonetary search time costs associated with making a purchase. Imposing milder sanctions for possessing smaller quantities may "bribe" users to purchase smaller quantities. The rate of purchase may increase, but it need not increase enough to offset the decreased purchase size.

The consumption minimizing policy is derived. It calls for no punishment at all for quantities below some threshold, so it may not

be politically feasible. The consumption minimizing policy within a restricted class of policies that might be politically feasible is also derived. Comparing various policies suggests that under ideal conditions implementing even the latter policy could significantly reduce consumption.

Conditions are not likely to be ideal in the real world, however, so the model is not well suited to quantitatively specifying what the punishment policy should be. It does make the point though that punishment policies which are increasing functions of the quantity possessed at the time of arrest may have desirable properties.

The chapter also briefly explores some of the possible benefits of having different punishments for repeat offenders.

The general problem of finding appropriate sanctions is both difficult and important. Chapter 5 merely scratches the surface by calling attention to one unexpected benefit of letting "the punishment fit the crime" by making the punishment an increasing function of the quantity possessed.

8.1.6 Chapter 6

Chapter 6 develops two models of how AIDS will affect the number of IV drug users. The first is a simple population model; the second is an open population version of a model developed by Kaplan.[1] Both come to essentially the same conclusion. If all other things (such as enforcement policy, the availability and price of injectable drugs, etc...) remain equal, in the long run AIDS will substantially reduce the number of IV drug users, perhaps by 50% or more.

There are at least two directions for further work in this area. The first involves obtaining better data. Currently relatively little solid data is available on the nature of needle sharing and risky sex practices of drug users either before or after the advent of AIDS. Better data would support better estimates of the parameters and hence allow for more accurate projections.

Second, the analysis in Chapter 6 focuses on steady state results. Clearly it would also be valuable to predict how the number of IV drug users will change over time. Some transient analysis can be done with the existing models. A more thorough analysis, however, would require developing more detailed models, for example by dividing the population by region, explicitly modeling sexual transmission, and distinguishing between different forms of needle sharing.

---

[1]Kaplan, 1989.

8.1.7 Chapter 7

Chapter 7 steps back and analyzes the industry level supply and demand curves for illicit drugs. Section 7.1 argues that the industry supply curve is downward sloping; the larger the market is, the less expensive it is per unit to provide the drugs. One reason for this is that when the market is large enforcement and punishment resources are spread thin. Another is that it is easier to make connections when there are many participants in the market. This directly reduces search time costs, and it increases competitive pressure, which squeezes out excess profits and inefficient business practices.

A consequence of a downward sloping supply curve is that there can be multiple, stable market equilibria. Thus, a downward sloping supply curve can explain both the high consumption at relatively low prices observed today and the relatively low consumption at higher prices of twenty years ago, without assuming there have been gross shifts in either demand or supply.

The most dramatic implication of the multiple equilibria model is that if society is at a high-volume, low price equilibrium, then increasing enforcement is unlikely to have much effect on consumption. Enforcement is much more likely to be effective at a low volume equilibrium. In contrast, demand reduction efforts, such as prevention programs, are likely to be particularly effective when the market is at a high volume equilibrium. The multiple equilibria model also counsels against imposing stiff minimum sentences.

Several researchers have speculated that the price elasticity of demand for drugs is probably low in the short run, but relatively high in the long run. The reason for this is that, because of addiction, demand at any time is a function of the quantity consumed in the past. Section 7.4 proposes a single functional form for the demand curve that yields both a low short run and a high long run price elasticity of demand.

A downward sloping supply curve and this type of demand curve both imply that, at least in the long run, the quantity consumed may be highly sensitive to exogenous changes. This instability suggests that one should be cautious when making predictions about untried polices by extrapolating from previous experiences.

8.2 Overall Conclusions

This final section makes a few general observations that do not belong under any single chapter.

8.2.1 The Drug System is Apparently Nonlinear and Unstable

Several of the models suggest that the drug system is nonlinear and unstable in the sense that small changes in one parameter may lead to large changes in another. The clearest example of this is the multiple equilibria model discussed in Chapter 7. It suggests that the market can be at a point such that even the smallest shock would push the market either to a low volume equilibrium or to a high volume equilibrium, depending on the nature of the shock.

The possibility that the long run price elasticity of demand is high (also discussed in Chapter 7) does not introduce such dramatic instability, but it does suggest that consumption may be more sensitive to exogenous changes than is generally assumed. Likewise, the analysis in Chapter 3 disputes the notion that interdiction is futile because retail prices, and hence consumption, cannot be significantly influenced by even a successful interdiction program.

The positive feedback effect (discussed in Chapter 4) implies that enforcement's effect on the size of a local market is not proportional to the level of resources allocated. Even a small shock can collapse a market that has already been reduced to its minimum viable size.

Of course systems that are sensitive at some points must be relatively insensitive at others. For example, the balloon model suggests that applying half the enforcement pressure required to collapse a market will only marginally reduce its size. Likewise, the multiple equilibria model suggest that when the market is at a high volume equilibrium, increasing enforcement is unlikely to have a dramatic effect on consumption.

A related issue has to do with how enforcement efforts should be allocated. If one believes the drug system is a stable, insensitive monolith, one might be tempted to attack on all sides to whittle away at the edges. Calls to attack all sides of the drug problem are common.[2]

This thesis (particularly Chapter 4) suggests taking quite the opposite approach, namely, "Go for the weak link." After all, enforcement need only eliminate one link between the farmers'

---

[2]For example, Guy (1990, p.9) states that President Bush's challenge to enforcement authorities is "to attack on all sides."

fields in the source country and the street corner in the United States to break the distribution chain. Enforcement is not obligated to attack every link; rather, it is the drug system which is obligated to defend every link.

An extreme example makes the point. If interdiction were 100% effective, then local street level enforcement against cocaine and heroin would be essentially superfluous. In general, there may be advantages to focusing efforts on those parameters to which the overall market is most sensitive rather than spreading efforts thin.

8.2.2 Enforcement Cannot "Solve" the Drug Problem

Most people initially seem to see the drug problem as a criminal justice issue and to believe that more or better enforcement is the proper response. Then as they become more familiar with the complexity of the issue, the limitations of enforcement, and its costs, they no longer see enforcement as a way to "solve" the drug problem. This has certainly been the case for me. Others, including respected researchers,[3] police and drug enforcement agents,[4] and even the John Lawn, the director of the Drug Enforcement Administration,[5] have come to the same conclusion.

The threat of enforcement may be able to suppress consensual crime when the market is small, but the larger the market is, the harder it is to credibly threaten stiff sanctions. This distinction between enforcement's effects on small and large markets arose with both the balloon model and the multiple equilibria model. They suggest that even substantial increases in enforcement are unlikely to greatly affect consumption. The history of enforcement in the 1980's supports this. The inputs to the criminal justice system grew substantially as did the number of arrests, convictions, and sentences for drug related offenses, but the quantity of drugs consumed appears to have increased not decreased.

Of course both the balloon model and the multiple equilibria model suggest that if enforcement were sufficiently severe, it could

---

[3]E.g. Kleiman, 1988b.

[4]Reinhold (1989) reports that police in many cities are beginning to stress prevention and quotes the Congressional testimony of Police Chief Charles A. Gruber of Schreveport, LA, "We understand that law enforcement is not the solution to the problem of drugs in our society."

[5]Lieber (1986, p.44) quotes Lawn's testimony before a Senate Committee "that when he joined the DEA, in 1982, he believed that 'with sufficient resources the drug problem could be addressed and solved with law enforcement alone.' He had come to understand, however, that 'the problem is greater than law enforcement is able to cope with.'"

make the market collapse (fall back to a low volume equilibrium). The key question is how Draconian enforcement measures would have to be to collapse the national market. Given recent history, one could be forgiven for being skeptical that such pressure can be brought to bear.

Even if enforcement can not "solve" the drug problem, it may be able to help by "holding the line" until demand decreases. Keeping drugs illegal, expensive, and relatively unavailable may be justified on the grounds that it holds down recruitment of new users. In a sense enforcement is doing all one can expect it to. After all, it makes common, processed agricultural products literally worth more than their weight in gold. In the long run demand reduction may have a better chance of "solving" the drug problem.

Demand reduction efforts are frequently divided into two categories: treatment and prevention. Both are important, but there is one little-realized reason for favoring prevention. Most people do not begin using on their own; they are introduced to drugs by a friend. Furthermore, it is thought that the majority of such initiations are done early in the "experienced" user's career.[6] Hence preventing one person from becoming a heavy user may do more to restrain future recruitment than helping one current heavy user quit.

That supply control is difficult does not imply that demand reduction will work, but there are at least two reasons for hope. First, demand reduction has not received as much attention or as many resources as supply control; so perhaps demand reduction has not worked better simply because it has not been given a chance. Second, for most drugs and most demographic groups, the prevalence of use has been declining substantially,[7] especially among young people.[8] Also, a recent poll found that 82% of 9th and 10th graders in Washington, D.C. surveyed "did not admire at all" a person who sold drugs; pimps were the only group less respected than drug dealers.[9] This suggests that the permissiveness of the 60's and 70's is being replaced by strong disapproval of drug use; and it may be that ultimately society's values are the primary determinant of drug use.

---

[6]Kaplan, 1983a.

[7]See Chapter 2's discussion of the National Household Survey.

[8]See Chapter 2's discussion of the High School Senior Survey.

[9]Reuter et al., 1990.

8.2.3 Enforcement Can "Manage" the Drug Problem

Suggesting that enforcement be thought of as a way to "hold the line" until demand falls does not imply that there is no room for improving enforcement policy. Quite the contrary, freeing enforcement from the obligation of trying to eradicate all drug use may allow it to more effectively "manage" the drug problem by taking steps to minimize the harm done by the drugs, the dealers, and the enforcement.

This view suggests, for example, using sanctions other than incarceration. Imprisonment is costly, not just to taxpayers, but also to the prisoner and to society. Incarcerating someone, in addition to its potential desirable effects of deterrence and isolation, also prevents the prisoner from working and hence in that sense is wasteful. Fines, on the other hand, simply transfer wealth, and community service actually produces something of value.

The argument against incarceration seems particularly strong for users and perhaps retailers. This coupled with the discussion in Chapter 5 suggests that minimum mandatory sentences and "getting tough" with casual users, especially first time offenders, may not be a good idea.

Differential punishment might be a way to take some of the violence out of drug dealing. For example, the standard punishment for smuggling could be 5 years in prison, smuggling while in possession of a firearm -- 10 years, and using it or brandishing it in a threatening way while smuggling -- 25 years. If the threat of the harsher sanctions exceeded the threat from other dealers that is avoided by carrying a firearm, then presumably smugglers would stop carrying firearms. One could similarly give extra punishment to dealers who employ children. These changes may not reduce the quantity consumed, but they might reduce the social cost associated with that level of consumption.

Recognizing that societal costs are not always proportional to the quantity consumed opens up other possibilities, for example, in source country control. The United States tries to eliminate drugs at their source. Occasionally these efforts are successful, but to date production has always quickly shifted to another country because it is easy and profitable to produce drugs, and there is no shortage of places to grow them.

A warped board metaphor may be appropriate. It may be possible to eradicate production in one country (nail down the end of the board that is sticking up), but production will soon spring back elsewhere (the other end of the board will pop up). Running back and forth nailing down one end of the board only to watch the other

end pop up may be less constructive than deciding which end of the board is least undesirable to have up and leaving that end up. For example, it may be less undesirable to have heroin produced in Southeast Asia than in Mexico because the production-related corruption and violence would be farther away. If so, then it may be a good idea to reduce source country control measures in Southeast Asia.

Enforcement can also play a role in prevention. Retail enforcement that drives drug dealing and drug use underground may increase search time costs, reduce the visibility of drug use, and reduce the violence associated with street dealing. So even if enforcement on the whole cannot solve the drug problem, attacking certain aspects of it, such as open-air dealing, may be worthwhile. Likewise, although the cocaine market may be too big to collapse, vigorous enforcement could keep the use of less common drugs, such as "ice", from becoming widespread.

In conclusion, trying to reduce the harm caused by the drug problem may be a more useful objective than trying to eradicate all use. If so, then there is probably room for more careful analysis and design, and mathematical modelling of the kind illustrated in this thesis may be able to contribute to that effort.

9: References

Adler, Patricia. 1985. *Wheeling and Dealing: An Ethnography of an Upper-Level Drug Dealing and Smuggling Community*, New York: Columbia University Press.

Alter, Jonathan and Mark Starr. 1990. "Race and Hype in a Divided City." *Newsweek* January 22, pp.21-22.

Anderson, R.M. and R.M. May. 1982. "Directly Transmitted Infectious Diseases: Control by Vaccination." *Science* Vol. 215, pp.1053-1060.

Bacchetti, P. and A.R. Moss. 1989. "Incubation Period of AIDS in San Francisco." *Nature* Vol. 338, pp.251-253.

Barnes, Deborah M. 1988a. "Drugs: Running the Numbers." *Science* Vol. 240, June 24, pp.1729-1731.

_____. 1988b. "The Biological Tangle of Addiction." *Science* Vol. 241, July 22, pp.415-417.

Barnett, Arnold. 1988. "Drug Crackdowns and Crime Rates: A Comment on the Kleiman Report." in *Street-Level Drug Enforcement: Examining the Issues*, Marcia R. Chaiken (ed.), National Institute of Justice.

Berke, Richard L. 1989. "Capital's Government Denounced By Bennett, Opening Drug Drive." *The New York Times* April 11, p.A1.

Besharov, Douglas J. 1989. "Cracked-Up Kids--Right from the Start." *The Washington Post National Weekly Edition* September 11-17, p.24.

Booth, William. 1988. "War Breaks Out Over Drug Research Agency." *Science* Vol. 241, August 5, pp.648-649.

Bouza, Anthony V. 1988. "Evaluating Street-Level Drug Enforcement." in *Street-Level Drug Enforcement: Examining the Issues*, Marcia R. Chaiken (ed.), National Institute of Justice.

Boyd, Gerald, M. 1989. "Bush, Citing Cost, Says Drug War Will Focus Largely on Education." *The New York Times* January 27.

Branigin, William. 1989. "A Stinging Indictment of High-Level Mexican Corruption." *The Washington Post National Weekly Edition* February 13-19, p.16.

Brown, George F. and Lester P. Silverman. 1974. "The Retail Price of Heroin: Estimation and Applications." *Journal of the American Statistical Associations* Vol. 69, September.

Buning, E.C., C. Hartgers, G. van Santen, A. Verster, and R.A. Coutipho. 1988. "A First Evaluation of the Needle/Syringe Exchange in Amsterdam, Holland." in *The Global Impact of AIDS*, eds. A.F. Fleming, M. Carballo, D.W. FitzSimons, M.R. Bailey, and J. Mann, New York: Alan R. Liss, Inc., pp.369-373.

Bureau of Justice Statistics. 1988. "Drug Law Violators 1980-86." Bureau of Justice Statistics Special Report, NCJ-111763.

_____. 1989. "Prisoners in 1988." Bureau of Justice Statistics Bulletin, NCJ-116315.

Bureau of Tobacco, Alcohol, and Firearms. "Summary Statistics on Distilled Spirits, Wine, Beer, Tobacco, Firearms, Enforcement, Taxes." Fiscal Years 1975, 1976, 1977, 1978, 1979, 1980-81, 1981-82 ATF P 1323.1, Transitional Quarter Statistics July 1-Sept.30, 1976 Distilled Spirits, Wine, Beer, Tobacco, Firearms, Enforcement, Taxes, ATF P 1323.2, and October 1982 - September 1988 ATF 0 5700.3 Industry Statistical Reports.

Bureau of Tobacco, Alcohol, and Firearms, 27 CFR Ch. 1, Section 285, April 1, 1988.

Burke, Kevin M. 1988. "Comments on Street-Level Drug Enforcement." in *Street-Level Drug Enforcement: Examining the Issues*, Marcia R. Chaiken (ed.), National Institute of Justice.

Canellos, Peter S. 1990. "From Here to LA, Search Policies Stir Debate." *The Boston Globe* January 15, p.1.

Canon, Lou. 1989. "A Supply-side Approach to the Drug War." *The Boston Globe* August 28, p.A15.

Centers for Disease Control. January 1990. *HIV/AIDS Surveillance Report*, Atlanta, GA.

_____. April 1990. *HIV/AIDS Surveillance Report*, Atlanta, GA.

Chaikan, Marcia R. (ed.). 1988. *Street-Level Drug Enforcement: Examining the Issues*. National Institute of Justice.

Committee on AIDS Research and the Behavioral, Social, and Statistical Sciences, National Research Council. 1989. *AIDS: Sexual Behavior and Intravenous Drug Use*, eds. C.F. Turner, H.G. Miller, and L.E. Moses, Washington, D.C.: National Academy Press.

Conners, Edward F., III. 1989. "Hazardous Chemicals From Clandestine Labs Pose Threat to Law Enforcement." *The Narc Officer* September, pp.25-29.

Controlled Substances Act, Chapter 13, Subchapter I, Part D, Section 841.

Daley, Suzanne and Michael Freitag. 1990. "Wrong Place at the Wrong Time: Stray Bullets Claim More Victims." *The New York Times* January 14, p.1.

Des Jarlais, D. 1988. "Current Epidemiology of AIDS Among IV Drug Users in New York City." in *Problems of Drug Dependence 1988*, National Institute on Drug Abuse Research Monograph No. 90. DHHS Publication No. (ADM) 89-1605, Washington, D.C.: U.S. Government Printing Office, pp.311-313.

Des Jarlais, D., S.R. Friedman, J.L. Sotheran, and R. Stoneburner. 1988. "The Sharing of Drug Injection Equipment and the AIDS Epidemic in New York City: The First Decade." in *Needle Sharing Among Intravenous Drug Abusers: National and International Perspectives*, eds. R.J. Battjes and R.W. Pickens, National Institute on Drug Abuse Research Monograph No. 80. Washington, D.C.: U.S. Government Printing Office, pp.160-175.

Des Jarlais, D.C., S.R. Friedman, and R.L. Stoneburner. 1988. "HIV Infection and Intravenous Drug Use: Critical Issues in Transmission Dynamics, Infection Outcomes, and Prevention." *Reviews of Infectious Diseases* Vol. 10, pp.151-158.

Des Jarlais, D. and D. Hunt. 1988. "AIDS and Intravenous Drug Use." National Institute of Justice AIDS Bulletin, NCJ 108620.

Drug Enforcement Administration. March 1985. "Drug Enforcement Administration: A Profile." Mimeo, prepared by Management Analysis Section, Office of the Controller.

_____. Various years. Office of Intelligence Illicit Drug Wholesale and Retail Price Reports. September, 1984; October-December, 1988; and April-June, 1989.

du Pont, Pierre S., IV. 1985. "Expanding Sentencing Options: A Governor's Perspective." National Institute of Justice Research in Brief, NCJ 96335.

Economic Report of the President Transmitted to the Congress: February, 1990. Washington, D.C.: United States Government Printing Office.

Feldman, H.W. and P. Biernacki. 1988. "The Ethnography of Needle Sharing Among Intravenous Drug Users and Implications for Public Policies and Intervention Strategies." in *Needle Sharing Among Intravenous Drug Abusers: National and International Perspectives*, eds. R.J. Battjes and R.W. Pickens, National Institute on Drug Abuse Research Monograph No. 80. Washington, D.C.: U.S. Government Printing Office, pp.29-39.

Friedland, G.H., C. Harris, C. Butkus-Small, D. Shine, B. Moll, W. Darrow, and R.S. Klein. 1989. "Intravenous Drug Abusers and the Acquired Immunodeficiency Syndrome (AIDS)." in *AIDS and IV Drug Abusers: Current Perspectives*, eds. R.P Galea, B.F. Lewis, and L.A. Baker, National Health Publishing, pp. 75-86.

Friedman, S., D. Des Jarlais, and J. Sotheran. 1989. "AIDS Health Education for Intravenous Drug Users." in *AIDS and IV Drug Abusers: Current Perspectives*, eds. R.P Galea, E.F. Lewis, and L.A. Baker, National Health Publishing, pp. 199-214.

Friedman, S., D. Des Jarlais, J. Sotheran, J. Garber, H. Cohen, and D. Smith. 1987. "AIDS and Self-Organization Among Intravenous Drug Users." *International Journal of the Addictions* Vol. 22, No. 3, pp.201-219.

Garreau, Joel. 1989. "Washington's Underworld Entrepreneurs: Applying Free-Market Analysis to the Drug Trade in Our

Nation's Capital." *The Washington Post National Weekly Edition* April 10-16, pp.10-11.

Guy, James P. 1990. "Call to Order." Presented at the International Narcotic Enforcement Officers Association Conference in Fort Lauderdale, Florida and reprinted in *The Narc Officer*, January, p.9.

Germani, Clara. 1988. "Combating the Drug Menace: Environment is Latest Coca Victim." *Christian Science Monitor* August 31.

Ginzburg, H.M. 1989. "Acquired Immune Deficiency Syndrome (AIDS) and Drug Abuse." in *AIDS and IV Drug Abusers: Current Perspectives*, eds. R.P Galea, B.F. Lewis, and L.A. Baker, National Health Publishing, pp. 61-74.

Gottfredson, Michael and Travis Hirschi. 1989. "If You Want a Successful Drug War, Take It to the Children." *The Washington Post National Weekly Edition* September 18-24, pp.23-24.

Gropper, Bernard, A. 1985. "Probing the Links Between Drugs and Crime." National Institute of Justice Research in Brief, NCJ 96668..

Harwood, H. J., D.M. Napolitano, P.L. Kristiansen, and J.J. Collins. 1984. *Economic Costs to Society of Alcohol and Drug Abuse and Mental Illness: 1980.* Research Triangle Park, NC: Research Triangle Institute.

Harwood, Richard. 1989. "Hyperbole Epidemic." *The Washington Post National Weekly Edition* October 9-15, p.29.

Hayeslip, David W., Jr. 1989. "Local-level Drug Enforcement: New Strategies." National Institute of Justice, NIJ Reports, No. 213, pp.2-7.

Hillsman, Sally T., Barry Mahoney, George F. Cole, and Bernard Auchter. 1987. "Fines as Criminal Sanctions." National Institute of Justice Research in Brief, NCJ 106773.

Hohler, Bob. 1989. "Gangs Choking Hartford's Inner City." *The Boston Globe* November 11, p.25.

Hosek, Linda and Paul Overberg. 1989. "Cracking the Hold of Ice." *The Rochester Democrat and Chronicle* Gannett News Service story, November.

Hundley, Tom. 1989. "Infants: A growing casualty of the drug epidemic." *The Chicago Tribune* October 16.

International Association of Chiefs of Police. 1989. "Reducing Crime by Reducing Drug Abuse: A Manual for Police Chiefs and Sheriffs."

International Drug Report. 1989a. "Record 20 Tons of Cocaine Seized in Los Angeles Warehouse." October/November, p.17.

_____. 1989b. "New Drug Rivals Crack in Lethality." Reprinted from the *San Francisco Chronicle* October/November, p.25.

Isikoff, M. 1989a. "Despite the Crack Epidemic There's Still Room for Heroin." *The Washington Post Weekly Edition* August 14-20, p.34.

_____. 1989b. "Just Say Outrage: A Recent Number Say Drugs Are America's No.1 Problem." *The Washington Post Weekly Edition* August 21-27, p.37.

_____. 1990. "Youths Deal a Snub to Drugs: High School Survey Shows Use Declining." *The Washington Post National Weekly Edition* February 19-25, p.38.

Jacobs, Sally. 1989. "Captives of Crack." *The Boston Globe* December 13, p.1.

Jetmore, Capt. Larry F. 1989. "The War On Drugs. Are We Winning?" *The Narc Officer* September, pp.79-80.

Johnson, et al. 1985. *Taking Care of Business: The Economics of Crime by Heroin Abusers.* Lexington, MA: Lexington Books.

Kalbfleisch, J.D. and J.F. Lawless. 1988. "Estimating the Incubation Period for AIDS Patients." Letter to *Nature* Vol. 333, pp.504-505.

Kantrowitz, Barbara, Pat Wingert, Nonny de la Pena, Jeanne Gordon, and Tim Padgett. 1990. "The Crack Children." *Newsweek* February 12, pp.62-63.

Kaplan, E. H. 1989. "Needles That Kill: Modeling Human Immunodeficiency Virus Transmission via Shared Drug Injection Equipment in Shooting Galleries." *Reviews of Infectious Diseases* Vol. 11, No. 2, pp.289-298.

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

Kaplan, J. 1983b. "Heroin." National Institute of Justice Crime File Study Guide, NCJ 97225.

Keeney, Ralph L. and Howard Raiffa. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* New York: John Wiley and Sons.

Kelling, George L. 1988. "Police and Communities: the Quiet Revolution." National Institute of Justice Perspectives on Policing, No. 1.

Kelling, George L. and James K. Stewart. 1989. "Neighborhoods and Police: The Maintenance of Civil Authority." National Institute of Justice Perspectives on Policing, No. 10.

Kleiman, Mark A.R. 1988a. "Crackdowns: The Effects of Intensive Enforcement on Retail Heroin Dealing." in *Street-Level Drug Enforcement: Examining the Issues*, Marcia R. Chaiken (ed.), National Institute of Justice.

_____. 1988b. "Drug Abuse Control Policy." Testimony Before The Senate Judiciary Committee. John F. Kennedy School of Government Working Paper #88-01-12.

_____. 1989. *Marijuana: Costs of Abuse, Costs of Control.* Westport, CT: Greenwood Press.

Kleiman, M. A.R. and Mockler R. A. 1988. "AIDS and Heroin: Strategies for Control." Project Report for the Urban Institute, Washington, D.C.

Kleiman, Mark A.R. and Kerry D. Smith. 1989. "State and Local Drug Enforcement: In Search of a Strategy." John F. Kennedy School of Government Working Paper #89-01-06.

Kouri, James J. 1989. "The Latin American Drug Trade and Narco-Terrorism: One Workable Solution." in *The Narc Officer* September, p.81.

Levin, Gilbert, Edward B. Roberts, and Gary B. Hirsch. 1975. *The Persistent Poppy: A Computer-Aided Search for Heroin Policy.* Cambridge, MA: Ballinger Publishing Company.

Lieber, James. 1986. "Coping with Cocaine." *The Atlantic Monthly* January, pp.39-48.

Lisowski, William. 1988. "Analysis of Wholesale Cocaine Price Data." Appendix B of *Sealing the Borders: The Effects of Increased Military Participation in Drug Interdiction* by Peter Reuter, Gordon Crawford, and Jonathan Cave, The RAND Corporation, R-3594-USDP.

Lui, K.J., W.W. Darrow, and G.W. Rutherford III. 1988. "A Model-Based Estimate of the Mean Incubation Period for AIDS in Homosexual Men." *Science* Vol. 240, pp.1333-1335.

Malcolm, Andrew H. 1989. "The Spreading Web of Crack." *The New York Times* October 1-2, Sec. 1, p.1.

Manning, Peter. 1980. *The Narc's Game: Organizational and Informational Limits on Drug Law Enforcement.* Cambridge, MA: MIT Press.

Marmor, M., D. Des Jarlais, H. Cohen, et al. 1987. "Risk Factors for Human Immunodeficiency Virus Among Intravenous Drug Users in New York." *AIDS: An International Bimonthly Journal* Vol. 1, pp.39-44.

Marshall, Eliot. 1988a. "Flying Blind in the War on Drugs." *Science* Vol. 240, June 17, pp.1605-1607.

_____. 1988b. "A War on Drugs with Real Troops?" *Science* Vol. 241, July 1, pp.13-15.

_____. 1988c. "Drug Wars: Legalization Gets a Hearing." *Science* Vol. 241, September 2, pp.1157-1159.

Martz, Larry, Mark Starr, and Todd Barrett. 1990. "A Murderous Hoax." *Newsweek* January 22, pp.16-21.

Martz, Larry. 1990. "A Dirty Drug Secret." *Newsweek* February 19, pp.74-77.

Mata, A.G. and J.S. Jorquez. 1988. "Mexican-American Intravenous Drug Users' Needle-Sharing Practices: Implications for AIDS Prevention." in *Needle Sharing Among Intravenous Drug Abusers: National and International Perspectives*, eds. R.J. Battjes and R.W. Pickens, National Institute on Drug Abuse Research Monograph No. 80. Washington, D.C.: U.S. Government Printing Office, pp.40-53.

May, R.M. and R.M. Anderson. 1987. "Transmission dynamics of HIV infection." *Nature* Vol. 325, pp.137-42.

Medley, G.F., R.M. Anderson, D.R. Cox, and L. Billard. 1987. "Incubation Period of AIDS in Patients Infected Via Blood Transfusion." *Nature* Vol. 328, pp.719-721.

Mieczowski, Tom. 1989. "Understanding Life in the Crack Culture: The Investigative Utility of the Drug Use Forecasting System." National Institute of Justice Reports, No.217, November/December, pp.7-9.

Miller, Mark. 1990. "A Failed 'Test Case': Washington's Drug War." *Newsweek* January 29, pp.28-29.

Mills, James. 1986. *The Underground Empire: Where Crime and Governments Embrace*. New York: Dell.

Mitchell, T., and R. Bell. 1980. *Drug Interdiction Operations by the Coast Guard*. Center for Naval Analyses.

Moore, Mark H. 1973. "Policies to Achieve Discrimination on the Effective Price of Heroin." *The American Economic Review* Vol. 63, No. 2, pp.270-277.

_____. 1977. *Buy and Bust: The Effective Regulation of an Illicit Market in Heroin*. Lexington, MA: Lexington Books.

_____. 1986. "Drug Policy and Organized Crime." in *America's Habit: Drug Abuse, Drug Trafficking and Organized Crime*. The President's Commission on Organized, Washington, D.C.

Moore, Mark H. and Mark A.R. Kleiman. 1990. "The Police and Drugs." *The Narc Officer* January, pp.31-45.

Morganthau, Tom, Mark Miller, David A. Kaplan, Todd Barrett, and Lynda Wright. 1990. "Uncivil Liberties? Debating Whether Drug-War Tactics Are Eroding Consitutional Rights." *Newsweek* April 23, pp.18-20.

Morin, Richard. 1990. "Trashing the Myth of Earth Day: Pollution Isn't Atop the Issue Hit Parade." *The Washington Post National Weekly Edition* April 23-29, p.37.

Morrison, June, Allen D. Putt, and Robert Zmud. 1975. "Estimating Heroin Use in an Urban Area." *The International Journal of Addictions* Vol.10, No.4, pp.589-598.

Musto, David F. 1987. *The American Disease: Origins of Narcotic Control*, New York: Oxford University Press.

Mydans, Seth. 1989. "Powerful Arms of Drug War Arousing Concern for Rights." *The New York Times* October 16, p.A1.

National Drug Enforcement Policy Board. 1987. *National and International Drug Law Enforcement Strategy.*

_____. 1988. *Progress Report 1987.*

National Institute on Drug Abuse. 1982. *Guide to Drug Abuse Research Terminology*. Jack E. Nelson, Helen Wallenstein Pearson, Mollie Sayers, Thomas J. Glynn, (eds.), DHHS Publication No. (ADM) 82-1237.

National Narcotics Intelligence Consumers Committee (NNICC). 1988. *The NNICC Report 1987.*

_____. 1989. *The NNICC Report 1988.*

Navarro, Nick. 1989. "Florida Sheriffs Launch "Operation Rockpile" 2224 Suspects Arrested in Initial Crack Raid." *The Narc Officer* September, p.21.

_____. 1990. "Drug Task Force Operations Planning and Future Strategies." Remarks made at the International Narcotics Enforcement Offices Association Conference, Nov. 12-18, 1989, printed in *The Narc Officer* January, pp.15-17.

Newcombe, R. 1989. "Preventing the Spread of HIV Infection Among and From Injecting Drug Users in the U.K." *The International Journal of Addictions* Vol. 1, No. 2, September/October, pp.20-27.

Office of Technology Assessment. 1986. *The Border Was on Drugs.* Washington, D.C.

Pitt, David E. 1989. "Surge in New York Drug Arrests Sets Off Criminal-Justice Crisis." *The New York Times* April 4, p.B1.

Power, R.M. 1988. "The Influence of AIDS Upon Patterns of Intravenous Use--Syringe and Needle Sharing--Among Illicit Drug Users in Britain." in *Needle Sharing Among Intravenous Drug Abusers: National and International Perspectives*, eds. R.J. Battjes and R.W. Pickens, National Institute on Drug Abuse Research Monograph No. 80. Washington, D.C.: U.S. Government Printing Office, pp.75-87.

Power, Robert and Brian Wells. 1989. "Responding to Crack." *The International Journal on Drug Policy* Vol. 1, No. 2, September/October, pp.13-15.

Preble, Edward and John J. Casey, Jr. 1969. "Taking Care of Business--The Heroin User's Life on the Street." *The International Journal of the Addictions* Vol. 4, No. 1 (March), pp.1-24.

Reinhold, Robert. 1989. "Police, Hard Pressed in Drug War, Are Turning to Preventive Efforts." *The New York Times* December 28, p.1.

Raiffa, Howard. 1968. *Decision Analysis: Introductory Lectures on Choices under Uncertainty.* New York: Random House.

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

_____. 1984. "The (Continued) Vitality of Mythical Numbers" *The Public Interest* No. 75 (Spring), pp.135-147.

_____. 1987. in "Combating International Drug Cartels: Issues for U.S. Policy." U.S. Senate Caucus on International Narcotics Control, September, p.164.

_____. 1988. "Quantity Illusions and Paradoxes of Drug Interdiction: Federal Intervention Into Vice Policy." *Law and Contemporary Problems* Vol. 51, No. 1, pp. 233-252.

Reuter, Peter, Gordon Crawford, and Jonathan Cave. 1988. *Sealing the Borders: The Effects of Increased Military Participation in Drug Interdiction.* The RAND Corporation, R-3594-USDP.

Reuter, Peter and John Haaga. 1989. *The Organization of High-Level Drug Markets: An Exploratory Study.* The RAND Corporation, N-2830-NIJ.

Reuter, P., J. Haaga, P. Murphy, and A. Prasckac. 1988. "Drug Use and Drug Programs in the Washington Metropolitan Area." The RAND Corporation, R-3655-GWRC.

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

Reuter, Peter, Robert MacCoun, Patrick Murphy, Allan Abrahamsen, and Barbara Simon. 1990. *Money from Crime: A Study of the Economics of Drug Retailing.* The RAND Corporation, R-3894-RF, forthcoming.

Reuter, Peter and Jonathan B. Rubinstein. 1989. "Fact, Fancy, and Organized Crime." *The Public Interest* No.53, pp.45-67.

Robinson, Eugene. 1989. "Dispensing Justice From the Bunkers: Prosecuting Colombian Drug Lords Is Easier Said Than Done." *The Washington Post National Weekly Edition* September 11-17, p.16.

Robinson, John. 1989. "Bennett Says Users Share Blame for Drug Violence." *The Boston Globe* March 20.

Rothenberg, Randall. 1990. "Speaking Softly of Life's Dangers." *The New York Times* February 16, p.D1.

Ryan, Richard A. and Ken Miller. 1989. "Drug Could Replace Crack When it Invades Mainland, Officials Say." *The Rochester Democrat and Chronicle* Gannett News Service story, November.

Schwartz, John, Elizabeth Bradburn, and Carolyn Friday. 1989. "Using 'Spies to Win a War'." *Newsweek* November 6, pp. 56-57.

Sciolino, Elaine. 1990. "World Drug Crop Up Sharply in 1989 Despite U.S. Effort." *The New York Times* March 2, p.1.

Selwyn, P.A., C. Feiner, C.P. Cox, C. Lipshutz, and R.L. Cohen. 1989. "Knowledge About AIDS and High-Risk Behavior Among Intravenous Drug Users in New York City." in *AIDS and IV Drug Abusers: Current Perspectives*, eds. R.P Galea, B.F. Lewis, and L.A. Baker, National Health Publishing, pp. 215-227.

Singer, Max. 1971. "The Vitality of Mythical Numbers." *The Public Interest* No. 23, Spring.

Sparrow, Malcolm K. 1988. "Implementing Community Policing." National Institute of Justice Perspectives on Policing, No. 9.

Spencer, Bruce D. 1989. "On the Accuracy of Estimates of Numbers of Intravenous Drug Users." in *AIDS: Sexual Behavior and Intravenous Drug Use*, eds. C.F. Turner, H.G. Miller, and L.E. Moses, Washington, D.C.: National Academy Press.

Stewart, James K. 1988. "Attorney General Announces NIJ Drug Use Forecasting System." National Institute of Justice Research in Action, Reprinted from NIJ Reports/SNI 208.

Sullivan, Bernard R. 1988. "Hartford Police Department Study and Recommendations."

Systems Research Corporation. 1985. *Review of Customs Service Marine Interdiction Program.*

Thompson, Tracy. 1989. "Can Justice Be A Little Too Blind?" *The Washington Post National Weekly Edition* September 4-10, pp.31-32.

Time. 1990. "Heroin Comes Back." February 19, p.63.

Tirole, Jean. 1988. *The Theory of Industrial Organization.* Cambridge, MA: MIT Press.

U.S. Bureau of the Census. 1987. *Statistical Abstract of the United States: 1988.* (108th edition.) Washington, D.C.

U.S. Department of Health and Human Services. 1987a. *Acquired Immune Deficiency and Chemical Dependence.* DHHS Publication No. (ADM) 88-1513. Washington, D.C.: U.S. Government Printing Office.

_____. 1987b. *Trends in Drug Abuse Related Hospital Room Episodes and Medical Examiner Cases for Selected Drugs,* DHHS Publication No. (ADM) 87-1524. Washington, D.C.: U.S. Government Printing Office.

_____. 1988a. *Demographic Characteristics and Patterns of Drug Use of Clients Admitted to Drug Abuse Treatment Programs in Selected States: Trend Data 1979-1984.* developed by CSR Incorporated for the Division of Epidemiology and Statistical Analysis, National Institute on Drug Abuse.

_____. 1988b. *National Household Survey on Drug Abuse: Main Findings 1985.* DHHS Pub. No. (ADM) 88-1586. Washington, D.C.: U.S. Government Printing Office.

_____. 1988c. *Illicit Drug Use, Smoking, and Drinking by America's High School Students, College Students, and Young Adults: 1975-1987.* DHHS Publication No. (ADM) 89-1602. Washington, D.C.: U.S. Government Printing Office.

_____. 1988d. *Data from the Drug Abuse Warning Network (DAWN): Annual Data, 1987.* Series I, Number 7, DHHS Publication No. (ADM) 88-1584. Washington, D.C.: U.S. Government Printing Office.

_____. 1989a. "NIDA Capsules: High School Senior Drug Use: 1975-1988." CAP 23, Issued by the Press Office of the National Institute on Drug Abuse, Rockville, MD.

_____. 1989b. "NIDA Capsules: College Students Survey on Drug Abuse: 1980-1988." Issued by the Press Office of the National Institute on Drug Abuse, Rockville, MD.

_____. 1989c. "National Household Survey on Drug Abuse: Population Estimates 1988." DHHS Publication No. (ADM) 89-1636. Washington, D.C.: U.S. Government Printing Office.

_____. 1989d. *Data from the Drug Abuse Warning Network (DAWN): Annual Data, 1988.* Series I, Number 8, DHHS

Publication No. (ADM) 89-1634. Washington, D.C.: U.S. Government Printing Office.

U.S. Department of Justice. 1989. "Drug Trafficking: A Report to The President of the United States." August 3.

U.S. General Accounting Office. 1983. *Federal Drug Interdiction Efforts Need Strong Central Oversight.* Report GGD-83-52. Washington, D.C.: U.S. Government Printing Office.

____. 1985. *Coordination of Federal Drug Interdiction Efforts.* Report to the Chairman, Subcommittee on Government Information, Justice, and Agriculture, Committee on Government Operations. Report GAO-85-67. Washington, D.C.: U.S. Government Printing Office.

U.S. Public Health Service. 1986. *Coolfont Report: A PHS Plan for Prevention and Control of the AIDS Users.* Public Health Report 101:341-348. Washington, D.C.:U.S. Government Printing Office.

Varian, Hal R. 1984. *Microeconomic Analysis.* New York: W.W. Norton & Co.

Weld, William F. 1988. "Public Corruption Is Costing Us Too Much." *The Washington Post National Weekly Edition* May 2-8, pp.22-23.

The White House. 1989. *National Drug Control Strategy.* Washington D.C.: U.S. Government Printing Office, September.

Wiant, Jon A. 1985. "Narcotics in the Golden Triangle." *The Washington Quarterly* Fall, pp.125-140.

Wish, Eric D. 1987. "Drug Use Forecasting: New York 1984 to 1986." National Institute of Justice Research in Action.

Wish, Eric D., and Joyce Ann O'Neil. 1989. "Drug Use Forecasting (DUF) January to March 1989." National Institute of Justice Research in Action, NCJ 119517.

Woodbury, Richard. 1989. "Cult of the Red-Haired Devil: A Drug Buse Uncovers and Evil Brew of Satanism and Murder." *Time* April 24, p.30.

Woodley, Richard. 1971. *Dealer: Portrait of a Cocaine Merchant.* New York: Holt, Rinehart, and Winston.