UNIVERSAL
LIBRARY
OU_156003
UNIVERSAL
LIBRARY

OSMANIA UNIVERSITY LIBRARY

Call No. 330 / B16P Accession No. G1216

Author Bain, Joe S.

Title Pricing, distribution and empl
1948

This book should be returned on or before the date last marked below.

PRICING, DISTRIBUTION, AND EMPLOYMENT

Economics of an Enterprise System

PRICING, DISTRIBUTION, AND EMPLOYMENT

Economics of an Enterprise System

JOE S. BAIN

University of California

HENRY HOLT AND COMPANY, NEW YORK

COPYRIGHT, 1948, BY HENRY HOLT AND COMPANY, INC.

December, 1948

PRINTED IN THE UNITED STATES OF AMERICA

To

Bea, Jennifer, and Terry

PREFACE

This book is written primarily for use in the first collegiate upper-division course in economic theory. It presupposes that the student will have had some survey course in economics but not necessarily one which goes very far with formal economic analysis. As such, I suppose it might be characterized as either "elementary" or "intermediate" economic theory. Recognizing that the content of the prerequisite survey course varies from place to place, I have begun on a rather elementary level and carried the subject as far as seemed desirable for undergraduate teaching.

As the student is introduced to economic theory, there is initially a serious question as to what should be emphasized. From one standpoint economic theory may be viewed as a system of formal reasoning that supplies certain analytical tools which can be applied to the solution of numerous practical problems. Correspondingly, the basic theory offering has often been constructed primarily as a "tool-box" course, emphasizing a technical training for the future economist. But economic theory may more broadly be viewed as a system of thought which offers an explanation of how the economy works and an evaluation of the tendencies that this system of thought discovers. The subject material is construed in the latter sense in this volume. My primary purpose is to explain to the student in some detail what theory tells him of the working of his econ-

omy and of the reasons this economy behaves as it does. Since the bulk of undergraduate "majors" in economics or business administration are obviously not going to become professional economists, it seems appropriate thus to reduce the emphasis on formal training and to focus attention mainly on the meaningful propositions which economic theory has developed.

Because they have been written with this emphasis, the following chapters do not present an especially intensive or advanced treatment of the formal apparatus of economic theory. The basic concepts and solutions are presented and discussed, but mainly as a means of showing what economic theory has to say and in general how it arrives at its conclusions. On the other hand, the work may be somewhat more comprehensive of the content of economic theory and of its implications than some more strictly formal treatments. The discussion of pricing and distribution, for example, is not limited to a consideration of the firm and industry (the particular equilibrium) but extends to cover a number of aspects of the interrelated behavior of all sectors of the economy (the general equilibrium). Thus pricing is considered not only as an individual industry problem but as a general phenomenon affecting the level of employment and the allocation of resources among uses.

The content of the volume includes a theory of price, in particular and general equilibrium aspects, a theory of the distribution of income, and a theory of the level of income and employment. Some of these matters have perhaps been more thoroughly explored than others. For example, the theory of employment is given only a synoptic treatment in connection with the discussion of capital and interest, and a course devoted primarily to income and employment should certainly seek much additional reading material. In general, however, I have attempted to prepare a primer covering a good deal of the content of modern economic theory. Certain exceptions to this may be noted. The theory of consumer choice is not developed but only referred to as an explanation of the observed system of demands for goods. The theory of production is not treated in great detail but is brought into the discussion mainly in the explanation of varying costs and of the inter-substitution of factors by the firm. These omissions are deliberate; if the in-

structor desires to emphasize these matters, a great many standard treatments of them are available.

Little of the basic theoretical content of this volume is original. I have, however, felt free to apportion emphasis to various lines of analysis in accordance with their empirical relevance and their pedagogical value. Thus the analysis of price is dominantly an analysis of monopoly price and primarily of pricing in oligopolistic situations. This would seem appropriate in view of the current structure of the American economy. I have discussed oligopoly pricing at considerably greater length than is customary in theory textbooks, in the belief that we should emphasize the pricing situations of the real world. The treatment of income and employment runs entirely in terms of a sequence analysis; some Keynsian ideas are put forward in a period-analysis format in order to bring them perhaps a bit closer to experience.

The materials contained herein have been mostly covered in the first semester of a year's course on theory—that devoted to pricing and distribution—and have served as an introduction to a second semester on income and employment. With appropriate selection and supplementation, I hope that they may be useful under various curricular arrangements.

My sincere thanks go to Professor William Fellner for his extremely helpful advice and criticism on many points throughout the manuscript. And I am deeply indebted to a number of my colleagues who have used a mimeographed version of the first nine chapters in teaching and have made many useful suggestions.

J. S. B.

University of California, Berkeley
February 10, 1948

CONTENTS

1 THE FUNCTION OF ECONOMIC THEORY 1

2 THE DEMAND FOR COMMODITIES 10

The determinants of enterprise sales policies · The demands for the outputs of firms and industries · The demand for the output of an industry · The fundamental determinants of the negatively sloping demand curve · Elasticity of demand · Further complexities of demand analysis · The demands for individual sellers' outputs · Single-firm monopoly · Pure competition · Monopolistic competition · Oligopoly · Summary

3 THE PRODUCTION COSTS OF THE FIRM 61

The definition of "cost" · Definition of the "short period" and of fixed and variable costs · Short-run relation of cost to output · Cost variation in the long run

4 PRICE DETERMINATION IN PURE COMPETITION . . . 95

Price determination for various time intervals · The significance of pure competition · Short-run pricing by the firm in pure competition · The short-run price for the industry in pure competition · The

long-run price for an industry in pure competition · Normative properties of industry behavior in pure competition · Normative properties of an economy in pure competition

5 PRICING IN MONOPOLIZED MARKETS 136

The institutional setting of monopoly · Pricing by a single-firm monopolist · Monopoly price results and the general economic welfare · Aspects of the dynamics of monopoly price policy · The impact of monopoly on general welfare—further remarks · Empirical evidence of monopoly behavior · Single-firm monopoly in the public utility field

6 PRICING AND PRICE POLICY IN OLIGOPOLISTIC MARKETS . 176

Subcategories of differentiated oligopoly · Pricing with independent action by several sellers · Collusive and concurrent pricing in differentiated oligopoly · Market shares in differentiated oligopoly · Nonprice competition · Pricing results in differentiated oligopoly—summary · Pure oligopoly · Oligopoly and the working of the economy

7 THE EFFECTS OF CONCENTRATED BUYING 222

Pricing under simple monopsony—one buyer supplied by many sellers · A few buyers supplied by many sellers—collusive oligopsony · Bilateral monopoly and bilateral oligopoly · Monopsony and allocation—further remarks

8 MARKETS IN MONOPOLISTIC COMPETITION 240

Pricing in monopolistic competition · Nonprice competition among many sellers · Monopolistic competition—a summary

9 THE PRICE SYSTEM FOR COMMODITIES 253

Resource allocation · Productive efficiency and income distribution · The size of selling costs · Progressiveness in the modern economy · Economic stability

10 THE DISTRIBUTION OF INCOME . 269

The problems of employment and income distribution · Productive factors and distributive shares · A general analysis of the distributive problem · Purchase of productive factors by the firm—pure competition · Purchase of productive factors by the industry · Purchases of productive factors by a competitive economy · The mutual determination of aggregate supply price and aggregate demand price · Summary

11 CAPITAL GOODS AND INVESTMENT 311

The character of capital · The demand for capital goods · Interest cost and the use of capital goods · Investable funds and general equilibrium · Dynamic change and net investment in capital goods · Technological progress and net investment · The timing of net investment · Net investment and gross investment · Investment in consumer finance

12 INTEREST, MONEY, AND EMPLOYMENT . 366

The supply of investable funds from saving · Cash balances, hoarding, and liquidity preference · The rate of interest for a single period · The bank rate of interest · The conditions for movement and stability of money income · The equilibrium of money income · Money price changes and the level of employment · Interest, income, and employ-

ment · The share of income earned as interest · Risk premiums and interest · Monopoly and monopsony as they affect capital goods and investable funds · The government debt, investment, and interest

13 THE THEORY OF FACTOR PRICING AS APPLIED TO RENTS AND WAGES 438

General considerations affecting wages and rents · Disparate submarkets for labor or land · The effect of commodity monopolies on wages and rents · The structure of labor markets · Monopsonistic buying of labor or other factors · Monopolistic selling of labor · Bilateral monopoly in labor markets · The general level and the structure of wages · Wage determination with fluctuating national income

14 THE PROFITS OF ENTERPRISE 476

The absence of profits in purely competitive equilibrium · Monopoly and monopsony profits · Dynamic change and profits · The recipients of profits from incorporated business · The distribution of income—ethical aspects

15 CONCLUSION . . . 488

Index 491

PRICING, DISTRIBUTION, AND EMPLOYMENT

Economics of an Enterprise System

CHAPTER 1

THE FUNCTION OF ECONOMIC THEORY

Theoretical economics, with which this book is concerned, should explain how an economy works—what sort of results it gives, and why it behaves as it does. An "economy," of course, is essentially a society of people engaged in their main occupation of making a living. People throughout the world are organized for the purpose of producing and distributing the things they want to use. All the persons in any country, or in the world, make up a sort of army constituted to accomplish the cooperative task of providing output for useful consumption. Their "economy" embraces organized effort in agriculture, mining, manufacture, transportation, marketing, merchandising, the service trades, and so on, as well as individual efforts in the arts and professions. The existence of an economy as a unified organization with a central purpose is especially clear in a fully socialized country. But it must also be recognized in a private-enterprise society like our own, where the pattern of economic effort emerges without central planning from the uncoordinated actions of a large number of people who are guided mainly by their pursuit of individual gain. The main difference is that in a fully socialized economy the ultimate economic accomplishment of the populace is largely premeditated, whereas in a free-enterprise economy it emerges as it will, automatically or by accident, and controlled mainly by the force of competition.

Whatever its form of organization, however, any economy performs in some measurable way, and its effectiveness may be appraised. There are several dimensions of the over-all accomplishment of an economy with which we are always closely concerned:

1. The amount of goods it produces, in the aggregate or per capita, each year or other time interval; in short, its *productivity* and, inferentially, its ability to employ available labor and resources.

2. The costs, in terms of human effort and physical resources used, which it incurs per unit of useful output it produces; roughly, its *efficiency*.

3. The way it distributes the goods it produces among its population; its *pattern of income distribution.*

4. The proportions in which it produces the various goods it makes, relative to consumer needs or desires for them—the *pattern of allocation of resources among alternative uses.*

5. The rapidity with which on the average the economy's output increases over time; its *progressiveness.*

6. The degree to which the output of the economy fluctuates over time; in short, the *stability* of the economy.

These are the principal dimensions of the material performance of any economy taken as a whole. With respect to each dimension, moreover, most people have an idea of what constitutes desirable performance. Large output and employment are preferred to small, at least up to some margin of overwork. (Substantial unemployment is almost uniformly disliked.) People also generally want low real costs [1] per unit of output, a composition of total output congruent with the pattern of consumer needs, and freedom from excessive economic instability. Although there may be more difference of opinion on the best sort of income distribution and the optimum degree of progressiveness, some majority sentiment on these matters can be established.

How a given economy behaves in each of the preceding respects—how close it comes to "ideal" or desirable behavior—

[1] By low "real costs" is meant a low expenditure of labor and resources per unit of output—low costs in terms of labor hours, machine hours, and materials used.

determines the level of material welfare of the people who are a part of it and depend upon it. It also determines whether public action will be considered necessary to improve the welfare-creating capacity of the economy. For these reasons, and also in the interest of better understanding of our society, it is important (1) to measure the performance of an economy in each of the several dimensions referred to, and (2) to explain why the economy gives the results that it does.

The measurement of the performance of the economy is essentially one of statistical enumeration and compilation, and is undertaken by many private and governmental agencies in our own society. Thus we have available manifold statistics of production, employment, cost, composition of output, and income distribution. From these statistics we can get an idea of how satisfactorily our economy has performed and where improvement might be desired. But measurement is not sufficient. If we are to appraise the significance of our findings or, more important, to consider means of improving observed results, it is essential that we understand their causes. We must know "what makes the economy tick" and why it behaves as it does. Only in this way can we learn whether the observed results, whether ideal or intolerable, are inevitable, accidental, or subject to modification with a modicum of effort.

Economic theory is largely concerned with explaining how an economy works and why it works as it does. In so doing it must explain certain basic forces which influence any economic activity, such as the character of consumers' desires for goods and the nature of productive techniques. It must also show the importance of the institutions or forms of organization through which economic activity is conducted—of the social planning authorities of a socialized economy or of the manifold institutions of a free-enterprise economy like our own.

The economic theory with which we are concerned here deals with the nature and behavior of a capitalist economy, with particular reference to the economy of the United States. It attempts to explain how a capitalist economy behaves, what sort of results it tends to give, and why it gives these results. As such it is in part pure scientific inquiry, but it is also an indispensable tool of our public policy.

The analysis of a free-enterprise economy, like any economic analysis, must begin with a consideration of the nature of human wants and the character of production techniques. To understand the functioning of any economy we must first establish the governing pattern of consumer desires for a variety of goods, and the basic relations of effort expended to output secured in producing various goods. The basic psychological and physical laws influencing all economies must be established. But our free-enterprise analysis must reflect with equal clarity the institutional characteristics of the modern capitalist economy which determine the sort of performance this economy gives.

The salient characteristic of modern capitalism is that it rests on a political system which recognizes a substantially universal exercise of the private ownership and operation of land, resources, and the physical equipment for production. Private individuals are thus able to own and control productive organizations of every type and to guide their operations in such a fashion as to enhance their individual gains. Such owners can secure the human labor necessary for production by hiring it at a wage, under a general system of free contract for human services. As a result, the activity of a capitalist economy is carried on mainly by a rather large number of private individual owners, or organizations of such owners in private business firms or corporations.

The activity of such private organizations is not directly guided by any central social responsibility to produce more or less, or this instead of that. Private enterprises in our system are expected to, and do, guide their courses of action in such a way as to make the largest possible individual profits [2]—they are profit-seeking enterprises. It follows that in our economy the prime decisions as to what to produce, how much to produce, how much labor and resources to employ, and what to charge the public are made by individual firms, and turn mainly on their effects on enterprise profits. Furthermore it is apparent that there is no central agency for coordinating individual profit-

[2] This is a good first approximation to the objective of enterprise activity; more extended consideration may cast doubt on the precise meaning of profit *maximization* and on whether firms always want the *largest* profit.

seeking decisions—they become coordinated only indirectly through the complicated mechanisms of markets for goods.

In our economy, moreover, individuals and firms do not produce primarily for use or for barter, but to sell for money. Individuals sell their labor for a money wage, and productive firms buy materials, equipment, and labor for money and sell their outputs for money. The pursuit of profit is precisely a pursuit of monetary profit—of a maximum difference between ultimate money expenditure and money income. Ours is a monetary exchange economy, where activity is immediately related to the flows of money expenditure seeking to purchase various goods and services.

In deciding, with an eye to profit, what and how much to produce and how many to employ, private business firms are guided by the general volume of money expenditure seeking goods, and by the tendency of this flow to seek specific goods and services. Production follows the aggregate money demand for goods and the composition of this demand. Similarly, individuals decide whether to work or not, where, and for how long at least partly in terms of the flow of money demands for the productive services they can supply.

As a result of the large number of independent decisions relating production or work to money demand, and as a result of the interactions of these decisions in the markets for goods and services (and for money itself), the over-all accomplishment of the capitalist economy emerges.

This is a brief description of the framework of action in a capitalist or free-enterprise economy. The task of economic theory is to explain what course this action takes and what results it ordinarily tends to have. This task, as economists are approaching it today, falls into two principal parts which, though related, are for expository purposes reasonably divisible.

One part of the work of theory is to explain what determines the size and course through time of the aggregate flow of money expenditure, production, and employment. In a money exchange economy like ours, productive activity and employment are elicited by the presence of effective money demand; the rate of output and employment at any time depends on the relation of

the flow of purchasing power to the general level of prices.[3] It follows that it is quite important to investigate the determinants of the rate of flow of money expenditure, and its relation to the price level, for the economy as an aggregate, in order to know under what circumstances a capitalist economy will give full employment, when it will give only partial employment, and according to what patterns over-all production and employment may be expected to fluctuate. An important division of economic theory is now directly concerned with this general problem of income, employment, and economic fluctuations; with showing how these things are jointly determined by individual behavior in disposing of cash income, by business decisions respecting investment, and by the character of our financial and banking institutions.

Another equally important task of economic theory is to explain the process by which in a capitalist economy the allocation of resources and the distribution of income are determined. There is very evidently a complex process of competition, bargaining, and selection of alternatives which eventually determines, at any level of employment, (1) how much of each good will be produced compared with others—*e.g.*, the ratio of automobiles to beefsteak to mechanical pencils; (2) what proportion of resources will be used in basic production and what in distribution and "selling"; (3) what shares of income will go to labor, to land and resource owners, to loan and investment capital, and to enterprise profits; and kindred matters. This process is embodied in general in the working of the *price system.* Commodities, services, resources, labor, and the use of money are bought and sold for prices under the general guide of the individual pursuit of profit or similar advantage. How much is produced of refrigerators as compared with roofing tile turns ultimately upon their relative prices; similarly the shares of income going to capital and labor are a price matter. In the nature and working of the price system lies the explanation of the various phases of resource allocation and income distribution in capitalism. The sector of economic theory which deals with

[3] That is, to the average level of "factor prices"—wages, rents, and interest—and to the average price of commodities produced by factors; the relation of commodity to factor prices is also important.

this explanation is generally known as price economics, or price analysis.

We will first be concerned in this volume with price analysis in its two main subdivisions—*commodity* price analysis and *factor* price analysis—which together attempt to explain the allocation of resources among uses and the distribution of income among various factors of production. Commodity price analysis is concerned generally with the explanation of how the profit-seeking and competitive activity of enterprises, and the reactions of the system of commodity prices to this activity, operate to allocate resources among uses or to determine what goods and services will be produced and in what proportions. It involves specific inquiry into the determination of the prices and outputs of individual commodities, and also of the quality or design of such goods, their costs of production, and the amount expended on selling costs. It also involves the analysis of the relations among the prices and outputs of all commodities and of the adjustments of these prices to costs of production. Such commodity price analysis may be conducted initially on the assumption that there is some given constant flow of money income or purchasing power demanding commodities as a whole, and also that there are given money prices for factors of production, so that the prices of those things which make up the costs of production are fixed. This procedure in effect enables us to ascertain how commodity prices, outputs, and so forth tend to adjust to any going level of money incomes and factor prices, and also enables us to observe the tendency of allocation of resources in any such given situation.

This analysis becomes somewhat more general when we relax the arbitrary assumption of given factor prices and suppose that, money income still being constant, the average level of money factor prices is also free to adjust to this flow of income. In this way we may ascertain not only the tendency in allocation of resources but also the adjustment of prices to costs and the determination of the ratio of profits to other factor earnings in any given income situation.

The second main step in price analysis is to admit that every factor price—wages, rent, interest—is free to adjust relative to commodity prices and to other factor prices, and to investigate

the determination of the relative prices of the various factors and the relative amounts of their employment. This will enable us to observe what fixes the relative incomes earned by labor, land, and capital and also to observe the simultaneous determination of all elements in the interrelated system of commodity and factor prices.

Both phases of this price analysis will be initially conducted on the assumption of a given fixed flow of money income, which amounts in effect, as we will see, to the assumption of a given level of employment. A final step in our analysis will be to recognize the flow of money income to be flexible and interrelated with the average level of prices. This leads to an inquiry into the determinants of the ratio between the flow of such income and the level of prices, and thus of the level of employment of labor and of other factors of production. Such analysis is evidently the key to the explanation of chronic or periodic unemployment in our economy, and also of cyclical and other fluctuations in employment and output.

The best starting point in the analysis of commodity and factor pricing is the operation of the individual firm or enterprise. In a capitalist economy such firms make the crucial decisions as to what to produce, how much to produce, what to charge, and so forth—and they make these decisions largely in terms of their effects on enterprise profits. The eventual behavior of prices and outputs for the economy as a whole emerges from these decisions and from their interactions through the competitive and bargaining process. We will therefore initially be concerned with the "economics of the firm," an investigation of the principles which govern certain strategic decisions made by the individual enterprise. This leads naturally into the analysis of the interactions of the decisions and resulting actions of related firms, in "industry" groups and for the economy as a whole, and of the manner in which the processes of competition and bargaining impose some systematic law of behavior on individual prices and outputs and on the allocation of resources among uses.

Our first concern is the economics of the individual firm. And in considering this, we turn initially to the notion of the *demand* for a commodity.

SUPPLEMENTARY READINGS

ALFRED MARSHALL, *Principles of Economics* (8th ed.), New York: The Macmillan Company, 1930, Book I.

LIONEL ROBBINS, *The Nature and Significance of Economic Science* (2d ed.), New York: The Macmillan Company, 1935.

JOSEPH SCHUMPETER, "The Nature and Necessity of a Price System," *Economic Reconstruction*, New York: Columbia University Press, 1934.

W. SOMBART, "Capitalism," *Encyclopedia of the Social Sciences*, vol. III.

THORSTEIN VEBLEN, *The Theory of Business Enterprise*, New York: Charles Scribner's Sons, 1932, Chaps. 1-4.

CHAPTER 2

THE DEMAND FOR COMMODITIES

THE DETERMINANTS OF ENTERPRISE SALES POLICIES

The explanation of commodity price formation begins with the decisions or choices made by the individual business enterprise. In a money exchange economy, such an enterprise operates by buying and selling. It purchases materials, equipment, land, and labor, combines them in a finished product, and sells this to a buyer. In so doing, it presumably attempts to "buy cheap and sell dear" or, more exactly, to maximize the difference between money income and money outgo over the time interval for which it makes advance calculations. This is fairly obvious. The question is—to what does the firm look in determining the precise course of action which will produce the maximum profit, or difference between receipts and expenditures? Our first task is to analyze the character of the controlling conditions to which the firm looks and which govern its activities in the pursuit of a profit.

At least five things are of immediate concern to such a firm—the product it will produce, the price at which it will sell, the quantity to be produced, the cost of producing it, and the amount of selling cost to be incurred in soliciting custom. The firm must decide what to produce, both generally and precisely—whether to produce cigarette lighters or fishing poles, and, if it is fishing poles, what type and what quality of poles. It must calculate what its chosen product will sell for, or the

alternative selling prices of each of a range of alternative products. It must decide how much to produce of the finally selected product. It must calculate how much this product will cost to produce, and how much to spend on advertising and other sales promotion. Out of these considerations it may decide what to produce, how much of it, what to spend, and what to charge to its customers.

As we consider this complex problem, two points are immediately evident: (1) that the various determinants of enterprise action, such as price, output, product, and cost, are not independent of one another, but rather interdependent, and (2) that none of these is single-valued or invariant—rather any of them, such as price, may assume different magnitudes as other determinants, such as output or product, are varied. In effect, each of the five determinants mentioned is a *variable* which depends upon other variables. It follows that the firm is not simply interested in price, output, product cost, selling cost, and product —which for convenience we may designate respectively p, q, c, s, and ϕ, but in the relationships of each of these variables to the others.[1] In fact, each variable is in a complex or multiple relationship to the others—thus price p depends on or is a function of quantity q, production cost c, selling cost s, and product ϕ. The problem may be, and often is, simplified, however, by considering the strategic relationships between certain pairs of the variables which determine the actions of firms.

The most significant relationships appear to be the following:

1. The relation of selling price p to output produced and sold q—the relation of sales receipts to the amount offered for sale. This can be calculated or estimated for any given product, selling cost, and production cost, and is ordinarily known as the *demand* relation. It is precisely the relation of p to q, when ϕ and s are held constant at chosen values. We will consider this relation at length in this chapter.

2. The relation of cost of production c to output q for any given product. (Price and selling cost will presumably not influence this relationship.) This measures the response of produc-

[1] Some arbitrary simplification is involved in regarding product as a single variable represented by a single symbol, ϕ, since product may be varied in numerous directions or dimensions.

tion cost to variations in output, subsuming a choice of production technique.

3. The relation of price p (or quantity q) to selling cost s, when the product and quantity (or price) are given. This measures the response of sales receipts or demand to variations in selling cost.

4. The relation of product ϕ to price (quantity being given) or quantity (price being given)—in effect the relation of product to demand—with given selling cost.

5. The relation of product ϕ to production cost c at any given output (selling cost and price presumably having no influence on this relationship).

In deciding what and how much to produce, what to sell for, and how much selling cost to incur, the firm will presumably take into account all these relationships, seeking a simultaneous solution which will yield the greatest possible profit. In succeeding chapters we shall explore the character of those calculations. For the moment we shall investigate in some detail the two leading relationships—the demand relation, of price to quantity sold (in this chapter) and the cost-output relation (in the next chapter).

THE DEMANDS FOR THE OUTPUTS OF FIRMS AND INDUSTRIES

Each firm has a primary interest in the demand for its own output; it wishes to know the various amounts it can sell per period of time at various alternative prices. The *demand schedule* for a firm should represent this information. It should show the number of units per month or per year which the firm can sell at each of a number of feasible alternative prices. Yet it is obvious that the demand for the firm's output, in this sense, is not an independent entity; the amounts the firm can sell depend on other things in addition to the price it charges. In an economy of many firms, the amounts which any one can sell at any price will obviously depend first on the type and design of product it makes, the extent to which it promotes its sale by advertising and other means, and the state of buyer preferences as between its output and the outputs of all other firms.

Let us take these things—product, selling costs, and buyer preferences—as given for the moment, and inquire further. The potential sales volume of the firm will then also depend upon the total volume of money purchasing power offered for all outputs and upon the prices of all other outputs. The greater the volume of money buying power, the greater the money price at which any firm will be able to sell a given amount, or the larger the amount it can sell at a given price. Further, the sales volume of any firm will also depend on the prices of other outputs which compete for the buyer's dollar—tending to be larger as the prices of other outputs are higher and smaller as other prices of other outputs are lower.

Looking at a single firm, it is then evident that in a given situation of buyer preferences, with a given product and selling cost, and if total purchasing power and the prices of all other outputs are given, this firm will have a determinate sales volume at each price it can charge, and thus a definite demand schedule relating quantity of sales to various alternative prices. Looking at all firms as a group, with given products, buyer preferences, and total purchasing power, it is clear that each in turn has a demand schedule which is dependent on the prices of all other sellers—so that in effect there is a family of demand schedules which are mutually interdependent.

To understand the character of the demand for any one firm's output, it is thus necessary, given total purchasing power and buyer preferences, to analyze (1) the extent to which this firm's sales volume responds to changes in its price, other prices being given, and (2) the extent to which other prices will respond to changes in the price of this firm, and the effect on this firm's demand of such induced or competitive price changes. The character of the interrelationships among the demands for various outputs and among various prices must be studied. As we approach such an investigation, however, it becomes evident that it is not expedient to consider each firm primarily in its relation to all other firms in the economy. Although the demand for the output of any firm is *at least slightly* influenced by the price charged by every other firm in the economy, the influence exerted by the prices of many such outputs is so small that it may for practical purposes be neglected. Thus the influence on

the demand for Camel cigarettes attributable to changes in the price of shoelaces is theoretically finite but obviously so small that it may be overlooked. For any one firm's output, moreover, there will ordinarily be a relatively small group of competing or substitute outputs, the prices of which have a fairly strong influence on the demand for this output. Firms are in effect found in competing groups, which we may call industries, within which the demands for the outputs of the members are closely interrelated, whereas the outputs outside of such a group will have prices with a small or insignificant influence on the demands for the given industry. Thus the demands for all brands of cigarettes are closely interrelated, as they are close substitutes, but the demand for any brand of cigarettes, or for cigarettes as a whole, will be little influenced by the prices of automobiles, building bricks, and bread. To analyze the character of the demand for the output of a single firm, therefore, it is necessary to recognize (1) the idea of an industry, or group of closely related or competitive firms, (2) the idea of a demand for the combined output of the industry—that is, a relation between the aggregated output of the several firms of the industry and the price or prices at which they sell, (3) the relationships among different industry demands, and (4) the character of the relation of the demand for a single firm's output to its industry demand and to the demands for the outputs of competing firms in the same industry.

A first step is to define an industry, and unfortunately this is not always a precise or neat notion. The primary basis for the delimitation of industries is the good or commodity. One group of sellers is set apart from another because they produce distinctly different goods, which fulfill different needs of buyers generally and are regarded by buyers as rather imperfect substitutes for each other. All kinds of goods are substitute sources of satisfaction, but as refrigerators and watches are not at all close substitutes to any buyer, they are different commodities, and their respective producers are in different industries. Thus initially we have as rough categories a steel industry, an aluminum industry, a breakfast-food industry, a ladies'-dress industry, and others to a number of several thousand. Within each industry is a group of producers, few or many, whose individual

outputs are very close substitutes for each other but whose aggregate output is a much more distant substitute for the outputs of other industries. Correspondingly the demands for the outputs of the member firms of an industry are closely interrelated, but they are more distantly related to the demands for outputs of other firms.

In ideal concept an industry is a group of sellers, the prices of whose products are very closely interrelated but have no significant relation to the price of any other seller or similar group of sellers. It is a separate focus for price determination not significantly related with any other focus for price determination. A group of sellers will be closely interrelated if they produce identical products (which are perfect substitutes for each other) and if they all sell to a common group of buyers. They will be isolated from other such groups if the substitution relation of their product or commodity to every other commodity is such that a change in their price will elicit no response from any other price sufficient to affect the demand for their product significantly—a condition we will more briefly characterize as the lack of any "significant" price interrelation with any other good. If a group of sellers fulfills these three conditions, they plainly constitute a distinct industry. Now if we could suppose that the total output of the whole economy were made up of a number of standard commodities each with no significant price interrelation with any other, and if the producers of each commodity produced identical products and sold in common to a single group of buyers, then the sellers of each commodity would constitute a distinct and fairly isolated industry. There would be no significant price interrelation between any pair of industries, although there would be a general interdependence of all industries. We could simply say, "An industry is a group of firms producing a single uniform commodity which has no significant price interrelation with another commodity and selling it in common to a single group of buyers," and we would find every firm in some such industry.

Unfortunately for purposes of analysis, actual industrial organization is not so simple. Three principal complications may be noted. In the first place, the outputs of the various firms producing a given sort of commodity are in most cases not identical

or perfect substitutes. There are instances where their outputs are in fact practically identical or homogeneous, as in the case of farmers producing wheat of a given grade and specification. In many instances, however, the outputs of firms producing the same sort of good are somewhat imperfect substitutes in the eyes of buyers, being *differentiated* one from another by design, quality, packaging, advertising, or direct sales promotion. Thus the various makes of automobiles are *differentiated products*, as are the various brands of cigarettes, the brands of soap, and so forth.[2] When this is the case, it is apparent that the output of one seller of cigarettes is related to another cigarette output in the same way in which it is related to another firm's automobile output—it is an imperfect substitute for either of them. The relationship of Camels to Chesterfields and the relationship of Camels to Buicks are different only in degree—Camels are a close substitute for Chesterfields and a distant substitute for Buicks.

This phenomenon requires alterations in the definition of an industry. Each seller of a slightly different good cannot be put in a separate "industry" if his price is in fact interrelated very closely with that of a number of close substitute outputs. Instead we may recognize an industry either as including identical products or as including a group of close substitute products with close price interrelationships. We obviously include in an industry a group of sellers whose products are either perfect or close substitutes, a change in the price of each of which can cause the price of the others to change enough to influence its own demand significantly. Where the number of close- or perfect-substitute outputs in a group is large enough that the price of no one seller affects the others' prices very much, we may still include in an industry the sellers of close-substitute products the demand for each of which is strongly influenced by concurrent price changes for the others. The industry includes a range of sellers of close-substitute products so defined; it excludes any seller the demand for whose output is not significantly influenced by the industry's price.

Using this definition, we can discover many clearly defined

[2] See Edward H. Chamberlin, *The Theory of Monopolistic Competition* (5th ed.; Cambridge, Mass., Harvard University Press, 1946), Chap. 4, for the principal original discussion of product differentiation.

industries in spite of product differentiation. But there is necessarily some imprecision and overlapping of industries, as where one or more firms will be fairly closely related to each of two otherwise independent groups of sellers. Thus firms A, B, and C may constitute an industry of close substitute products, and firms F, G, and H another distinct industry, but firms D and E may be closely related by substitutability of outputs to both groups. The phenomenon of overlapping or imprecisely limited industries requires careful analytical treatment. Recognizing this potential difficulty, however, we may define an industry as a group of firms producing either identical or close-substitute outputs, set apart from other firms the prices of whose outputs will not respond significantly to changes in the industry price.

Second, we shall occasionally find two groups of firms each of which qualifies in part as a separate industry under the preceding definition, but with some significant interrelation of the two industry prices—that is, there is a *close* substitution relation within each group, and further no seller in one group can alone significantly influence the price of any seller in the other group. But there may be a *more distant* but significant substitution relation between the two group outputs, so that a change in either group price somewhat affects the demands for the outputs of the other group, and may thereby elicit significant price reactions. When this occurs, the two industries are of course not strictly independent, but the respective groups of sellers may nevertheless be recognized as quasi-separate industries. The analysis of price determination in such cases should take account of the interaction of the related group prices.

The third major qualification in defining an industry stems from the fact that not all the firms producing a given range of close-substitute outputs will necessarily sell to a common group of buyers or market. The world market for most goods is broken up into continental submarkets by the force of transport cost, and further into national markets by political boundaries, tariffs, trade restrictions, and so forth. The European and American producers of many goods, for example, sell to largely exclusive groups of buyers. Within the United States, moreover, the sellers of many locally produced or hard-to-transport items supply local groups of buyers which are not reached by sellers

located in other states or regions. An industry, for analytical purposes, should include a group of sellers with close-substitute outputs which are sold in common to a single group of buyers. It should exclude sellers of the same good supplying an entirely different group of buyers. An industry thus has geographical or market limits as well as a commodity limit. Such geographical limits are seldom precise, and overlaps are very common. In practice we must be content to recognize an industry as including a group of sellers of close-substitute outputs which are sold in large part to a common group of buyers, and only in small part to buyers not supplied in common. Thus the automobile manufacturers of the United States constitute an industry for practical purposes, because they all offer the bulk of their outputs on a nationwide basis to all potential American buyers, whereas they ship only a small fraction of their outputs into the foreign market (thus overlapping a bit with the European automobile industry). In precise logic, an industry is a group of sellers of close-substitute outputs each of which offers his entire output to a common group of buyers. (It is this simplified conception of an industry which we will employ immediately below.) In practice, an industry is any workably close approximation to this logical ideal.[3]

An industry being thus defined, we may view every firm as being the member of some industry of one or more firms, with an output which is a close substitute for those of other firms in the same industry and is a more distant substitute for those of firms outside the industry. (Overlaps and in-between firms are of course allowed for.) For the aggregate output of each industry there is at any time some demand, or schedule of amounts the firms of the industry can sell at various common prices. The demand for the output of any firm will depend upon this industry demand and also upon the competitive relation between it and other firms in the same industry. We will therefore investigate in turn (1) the character of demands for the outputs of

[3] See Joe S. Bain, *Economics of the Pacific Coast Petroleum Industry* (Berkeley, Calif., University of California Bureau of Business and Economic Research, 1944), Part I, pp. 10-11, for a further discussion of the definition of an industry; also George J. Stigler, *The Theory of Price* (New York, The Macmillan Company, 1946), pp. 280-283.

industries, and (2) the character of the internal structure of industries, and how this influences the demand for individual firm outputs.

THE DEMAND FOR THE OUTPUT OF AN INDUSTRY

To investigate the properties of the demand for the output of a single industry, let us take the simple case of what we will assume to be a clearly defined industry of firms producing identical or homogeneous outputs and selling entirely to a common group of buyers. We will assume a single good—let us say gray (unbleached) cotton yarn—produced by a number of firms and sold entirely in a single market, that of the continental United States. The demand for gray cotton yarn will thus refer to the demand in this specific market alone. It must also refer to some specific time interval—let us say the amount to be bought in some certain month. We will further assume—perhaps somewhat inaccurately—that for no one substitute commodity will the price respond to changes in the price of cotton yarn enough to affect the demand for cotton yarn significantly, though of course such prices may change significantly for other reasons. This sets the cotton yarn industry clearly apart from every other and simplifies the case. Our problem is to analyze the demand for the gray cotton yarn industry of the United States, and for the United States market for a chosen time period of one month. What are the essential properties of such a demand?

The amount of a commodity which buyers will take in a given month will depend upon the choice pattern of these buyers as between this good and others, the volume of money purchasing power buying goods in general, the prices of other goods, and the price of the good in question. Each of these circumstances will influence the quantity of the good which buyers take. Let us center attention first, however, on the relation of the price of a good—gray cotton yarn—to the quantity of it which buyers will take. To do this, we shall suppose that buyers' tastes, total purchasing power, and the prices of all other goods are given and fixed at certain levels—either absolutely constant

or sufficiently invariant to have no perceptible effect on the demand for cotton yarn. (These being given, the goods produced by the industry in question will tend to sell a volume lying within some corresponding range—steam yachts may tend to sell 3 units per month, or gray cotton yarn in the range of from 650 to 700 million units.) And we shall investigate only the relation of cotton yarn prices to cotton yarn sales—the extent to which the number of units of cotton yarn bought would be influenced by variations in its price within this given situation.

It is initially evident that the amount bought will depend on the price charged and will probably become larger as the price becomes lower. *There is ordinarily an inverse relationship between sales volume and price.* At any one price, such as $1.25 per unit, a certain number of units, such as 672 million, may be bought. If the price were $1.30, fewer units would be bought; if it were $1.20, more units would be bought. At each specific alternative price, there should be a specific corresponding sales volume, which tends to become larger as price becomes lower. Such a relationship of price to sales volume would be observed not only for cotton yarn but for nearly every good.

This relationship of sales volume to price may be illustrated in a *demand schedule.* A demand schedule shows, for a given market and time interval, the volume of purchases which would take place at each of several alternative prices. Such a schedule might look as follows:

Price of yarn (per unit)	Sales volume of yarn (millions of units)
$1.30	661
1.29	662
1.28	663
1.27	665
1.26	668
1.25	672
1.24	676
1.23	680
1.22	684
1.21	688
1.20	692

The schedule could be extended to include any range of prices, but we should ordinarily be interested in a short range of prices which might be experienced within the chosen time interval. The schedule shows how many units all purchasers from the industry would buy in the given month if the price were alternatively each of those shown. Thus if the price were $1.28, buyers would take 663 million units per month; if *instead* it were $1.22, they would take 684 million units, etc. The schedule is drawn on the assumption of a given and unchanging purchasing power throughout the month and of a given state of buyers' wants.

To put our estimates in figures, however, we have assumed to be temporarily unchanging certain other things (in addition to income and consumer taste) which might influence the sales volume of cotton yarn. These include the prices of substitute yarns—rayon, nylon, and linen—and of any other substitute goods variations in the price of which would perceptibly influence the sales of cotton yarn. In effect, the demand schedule shows the *net* relationship of cotton yarn price to cotton yarn sales volume, assuming that all other things influencing cotton yarn sales are for the moment unchanging.

This idea may be made more precise by employing mathematical notation. In effect, we have before us several *variables* (a *variable* is a quantity which assumes various successive sizes). Two of these variables are the price of gray cotton yarn and the sales volume per month of this yarn in a given market. We may denote them as p_c (for price of cotton yarn) and q_c (for quantity of cotton yarn sales). Now p_c is related to (or is "a function of") q_c in a certain specific way. When p_c changes, q_c undergoes a definite corresponding change. We wish to know the exact relationship of the two variables. But q_c is *also* related to several other variables—the price of rayon yarn (p_r) the price of linen yarn (p_l), buyers' income (I), etc.—and changes also in response to changes in each of these. To get the *net* relationship of q_c to p_c—*i.e.*, to get the variation of cotton yarn sales in response to variations in cotton yarn price *alone*, and uninfluenced by variations in p_r, p_l, and I—we assume that p_r, p_l, and I are constant at given levels (that each is sufficiently invariant to have no perceptible effect on p_c), and *isolate* the effect of p_c

on q_c. It is this isolated effect that the demand schedule should show us.

This "*ceteris paribus*" (other-things-being-equal) demand schedule is a meaningful conception if in fact no other price nor income will change *in response* to changes in the cotton yarn price enough to influence the cotton yarn demand. Suppose, however, that some prices will so change, and will thereby affect the demand for cotton yarn. If this change follows no determinate pattern, no determinate independent relation of cotton yarn sales to price can be found. If, however, the other variables respond in a determinate and predictable fashion to cotton yarn prices, then a determinate demand schedule for cotton yarn can be defined on the assumption of predictable covariations in those variables. Such a demand schedule—a so-called *mutatis mutandis* schedule—is the alternative to the *ceteris paribus* schedule in such cases. For simplicity in the succeeding argument, however, we will view industry demand curves as being of the simpler *ceteris paribus* variety.

Statisticians have attempted to find the market demand schedule for various products, such as potatoes, steel ingots, and automobiles, and have arrived at numerical results generally showing inverse net relationships of price and quantity for given goods. They encounter difficulty, of course, in eliminating properly the effect of the "other variables" such as prices of competing products. Whatever the difficulties of statistical measurement, however, the idea of a demand schedule relating price and sales volume for a given good is evidently useful and valid.

The *demand schedule* for a good, in summary, shows the net relationship of the price of that good to the amount of it which buyers will take. Such a schedule may also be represented graphically on a pair of coordinate axes as in Figure 1. This graph presents the same information contained in the table on page 20. The price of cotton yarn is measured along the vertical axis and the quantity (bought) along the horizontal. There are eleven points for the eleven prices in the table, and each point is a *coordinate* showing the quantity from the table which corresponds to one price. Point A, for example, lies a vertical distance of 1.25 from the zero point on the price scale, and a horizontal distance of 672 from the zero point on the quantity scale; it

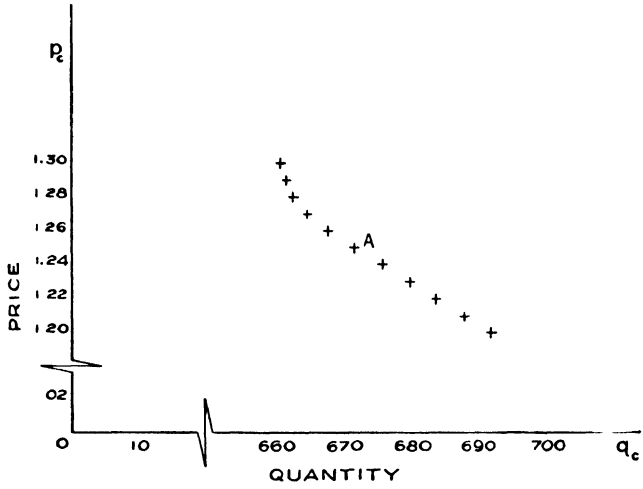

Figure 1

thus shows that the quantity 672 corresponds to (would be bought at) the price $1.25. Other points show similar relationships. The advantage of this graphic presentation is that when we observe the relation of one coordinate point to another we get a vivid idea of the way in which the quantity bought *responds* to changes in price. In the preceding example a moderate reaction is noted.

It is but a short step in graphic analysis to consider these successive coordinate points as connected up with a line, or, more conveniently, to suppose that we have a series of successive price changes each of which is indefinitely small, so that the succession of points makes up a practically continuous line. In this way we pass from a graph of successive discrete points, showing discrete quantities for discrete prices (Fig. 1), to a continuous line, showing the relation of quantity to price for every conceivable price. This line is called a *demand curve*; the one in Figure 2 is drawn from the same data shown in Figure 1. The demand line or curve (which we label *DD'*) shows the general relation of price to quantity bought for cotton yarn. Because it is a continuous line, it shows the quantity for each and every possible price down to the millionth or smaller fraction of a cent. To derive such a curve in practice, we necessarily

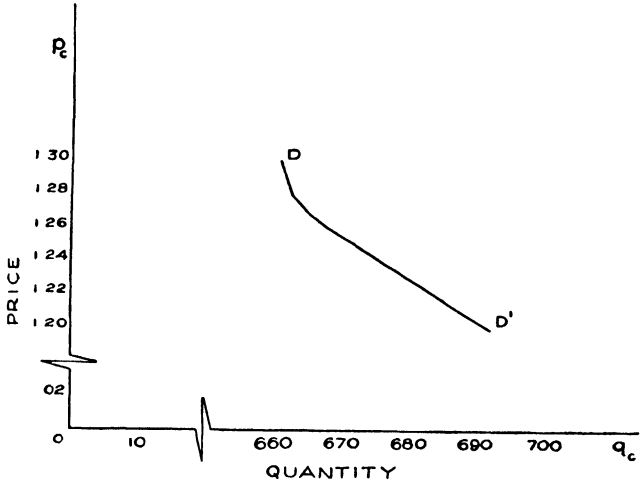

Figure 2

work with a number of discrete prices and interpolate between them, or "fit" a line to them, ordinarily by some statistical regression technique.

The student may well inquire what purpose there is in getting a continuous line for every conceivable shading of price. Admittedly, the businessman is interested only in a series of discrete prices, at perhaps one-cent—or even five-cent—intervals. The justification for setting up a continuous demand curve is twofold. First, a continuous line is more susceptible to analysis than a series of discrete points. Such a line, without in any significant way distorting the picture, enables us to think about pricing problems more effectively. Second, fitting such a line statistically is an expedient way of eliminating random deviations in the actual data which would otherwise confuse the problem.

The demand curve shown above, representing the net relationship of price to quantity bought of the output of an entire industry, is an *industry or market demand curve.* If the market should be supplied by a single seller—a monopolist—it would also be an individual seller's demand curve. Otherwise it is the common demand for the outputs of the several or many sellers who supply the market. In the example we have chosen, it is

the demand by all buyers in the United States for gray cotton yarn, and shows (under given circumstances) how much yarn all these buyers would take from all sellers at each possible alternative price within the range of the curve.

Such an industry demand curve has two properties of particular relevance to the analysis of pricing—its *position* and its *shape*. The *position* of a demand curve refers primarily to how far from the zero point it lies in the horizontal direction at any price—*i.e.*, to how large an amount is bought at any given price, or on the average at any of a number of prices. Suppose for a certain product we have a demand curve which always has the shape of a straight line with a 45° slope (if the price scale and quantity scale are constructed to given dimensions). For some given month and in given circumstances the demand curve for this product will look like line *a* in Figure 3. This line shows how much would be bought of the product at each possible price in the given circumstances of that month—with the given income, buyers' tastes, prices of other products, etc. But in a succeeding month the curve might lie at *b*, because perhaps of an increase in income, or an increased desire of buyers for the product, or an increased price of substitute products. In still another month the curve might shift to *c*, because of smaller income, decreased buyer desires, or lower prices of substitutes. The curves *a*, *b*, and *c* represent the market demand curve for a single product in different positions, or show how a demand curve may *shift* from position to position with changes in the surrounding circumstances. When a demand curve shifts, ordinarily a different quantity will be bought at every price than was bought before it shifted. It is not necessary, however, that a demand curve should retain the same shape as it shifts.

It will be inferred that when we employ a demand curve in analysis the effect of the changes of the price of a product on its own sales is shown by the shape of the curve—*i.e.*, by moving along a given curve—but that the effect of all other changes on the sales of the product are shown by shifts in the curve—by changes in its position. The position of the market demand curve is, of course, a dominant consideration in price determination, and the succeeding discussion of the shape of the curve should not be allowed to obscure this fact.

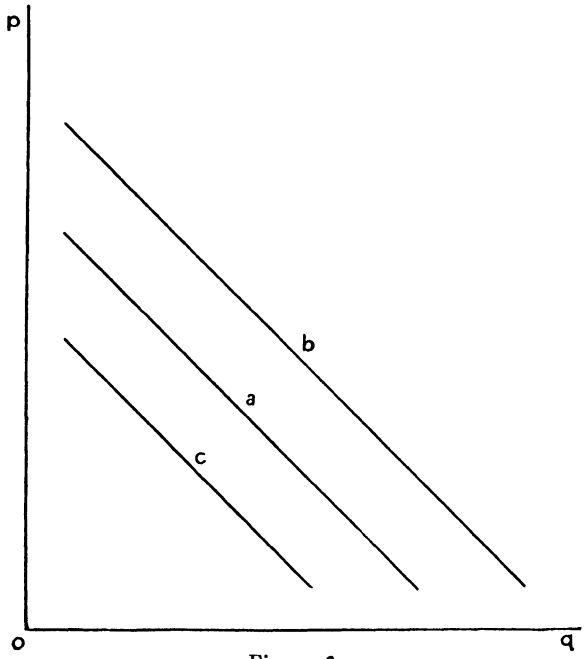

Figure 3

THE FUNDAMENTAL DETERMINANTS OF THE NEGATIVELY SLOPING DEMAND CURVE

The shape of a market demand curve involves the *direction* in which it slopes, the steepness of its slope, and characteristics of change of slope such as degrees and directions of curvature. Practically all market demand curves slope "downward to the right," or are *negatively* sloped—that is, quantity sold increases as price decreases (if other prices, buyers' tastes, and total purchasing power are unchanged), so that the price change is negative when the quantity change is positive, and vice versa. Some curves may slope steeply and others gradually, but nearly all of them presumably slope *negatively* at one rate or another.

Since the general principle that industry demand curves slope downward to the right is fundamental to further analysis, it may be well to consider the basis of the proposition. The assertion

that demand curves slope negatively rests mainly on observations of buyers' psychology. For goods bought for direct consumption, such as shoes, the argument runs somewhat as follows: The consumer ordinarily has limited means; he cannot acquire all he wants of every good. There is available to him an assortment of desirable goods, which are in general substitute sources of satisfaction, but the amounts of them the buyer can acquire are restricted by the limitation of his purchasing power. His problem is to apportion his spending among all available goods in such a way as to maximize the aggregate satisfaction he receives from his total purchases. Given the prices for all goods, therefore, he will try to juggle the relative desirability (to him) of various goods against their relative costliness until he arrives at that pattern of expenditures which seems to give him "the most for his money." Arriving at such a determinate spending pattern is made possible by the fact that as the consumer considers acquiring more and more of any one good (in place of other goods), additional increments to his purchases of that good become relatively less desirable; he can strike the best balance where the relative desirabilities of *the last increments* to his purchases of any two goods are in the same ratio as their prices. In short, he should bring into balance relative "marginal" desirability and relative price for all goods. Now if the price of any one good declines, the consumer will seek a new balance of desirability and relative price, and to do so he will substitute the cheapened good for other goods until this new balance is reached. This increases his purchases of the cheapened good. Furthermore, the reduction of the price of any good frees more purchasing power in the hands of users, thus stimulating their purchases of all goods, including those of the immediately cheapened good. A reduction in the price of any *consumer's good,* in sum, should tend to increase its sales. On the other hand, a rise in the price of such a good will ordinarily discourage its use, by pinching the buying power of users and by inducing them to substitute other goods for it.[4]

[4] In sum, the effect of a decline in the price of any good on its quantity of sales results from (1) substitution of this good for others, and (2) the effect

The preceding argument holds for consumer's goods, the purchase of which depends mainly on the satisfaction to be realized from these goods by their buyers. Many goods, however, are not bought by final users but are acquired by manufacturers for processing or use in production, or by middlemen for distribution. The purchase of these goods does not turn directly upon the satisfaction they could give to their buyers but upon the profit to be gained in producing with them or reselling them. Nevertheless, the demand curves for such *producer's goods* will also be negatively sloped—*i.e.*, the quantity purchased will increase with price reductions, and vice versa. This is because the demand for a producer's good is necessarily *derived* from the demand for some ultimate consumer's good, and the *derived demand* will ordinarily have the same general slope as the *primary demand.*

The demand for steel by automobile manufacturers, for example, is derived from the demand for automobiles by consumers. (The demand curve for automobiles is presumably sloped negatively for reasons mentioned above.) A price reduction in steel will stimulate the demand for steel if it induces the automobile manufacturers to make more automobiles, and it will do this so far as it leads to a reduction of automobile costs and prices, which in turn should stimulate automobile sales. A steel-price reduction will thus tend to increase steel purchases by automobile makers, although perhaps not by very much, by giving rise to lower automobile prices and increased automobile sales. (A steel-price increase should have the reverse effect.) A steel-price reduction may also stimulate steel use so far as it induces automobile manufacturers to substitute steel for other metals in making automobiles, or to use more steel per automobile. This tendency also causes the derived demand curve to slope in the same direction as the primary demand curve. But since the outlay for any one producer's good is ordinarily only a minor part of the cost of the final product in the production of which it is used, the industrial buyer of such a producer's good will

of the increased real purchasing power of consumers on the sales of all goods. See J. R. Hicks, *Value and Capital* (London, Oxford University Press, 1939), Part I, for an extended discussion of the theory of consumer choice and the derivation of demand curves.

ordinarily not react very strongly to moderate changes in its price.[5] As a general rule, therefore, the market demand curve for a producer's good—a derived demand—will be negatively sloped like the primary demand curve from which it derives, but the quantity of the good which buyers take will be less responsive to price changes than it is in the case of the primary good.[6]

The preceding support of the idea that industry demand curves are negatively sloped proceeds largely from general observations concerning buyers' psychology and behavior. The idea is fairly well verified by statistical measures of the market demand curves for many products.[7] If we choose any product and set out to find the statistical net relationship over a period of time between price charged and quantity sold, we will ordinarily find a negatively sloping demand curve.

It should also be noted that in constructing an industry demand curve, showing definite amounts which buyers as a whole will take at definite prices, we implicitly presuppose that in the market there are many small buyers, no one of which buys enough that he can perceptibly influence the market price. As a consequence, each buyer takes any going price as given and simply adjusts his purchases to this price—he does not try to drive price down by purchasing less or more. When all buyers behave in this way there emerges a definite purchase volume at each of several prices, with no bargaining by buyers. This is indeed the most common situation. The situation where buyers are few requires special treatment and will be dealt with in Chapter 7.

[5] The main exception to this would occur if one producer's good could be extensively substituted for another if its price fell enough. Aluminum is probably a case in point.

[6] Precisely, the elasticity of a derived demand for any good *A* depends on the elasticity of the primary demand (from which it is derived), the proportion of the cost of the primary good which is spent on good *A*, the substitution conditions between good *A* and the other goods or factors used in making the primary good, the supply conditions for substitute goods or factors, and the response of the price of the primary good to changes in its costs.

[7] See, for example, Henry Schultz, *The Theory and Measurement of Demand.* Chicago, University of Chicago Press, 1938.

ELASTICITY OF DEMAND

A general proposition concerning the shape of market demand curves, therefore, is that they all slope downward to the right, evidencing an inverse relationship between price and quantity. This is about all that various different market demand curves have in common, however. For although a series of price reductions for any good will ordinarily elicit a corresponding series of increases in quantity bought (tastes, purchasing power, and other prices remaining unchanged), the rate at which quantity responds to price may vary widely. A 10-percent reduction in the price of one good might bring about a 20-percent increase in its sales, whereas for another good a 10-percent price reduction might be accompanied by only a 2-percent increase in sales. The demand for the first good is much more responsive to price change than that for the second. Suppose that of good *A*, 100 pounds is bought at 10 cents and 120 pounds at 9 cents; but that of good *B*, 100 pounds is bought at 10 cents and 102 pounds at 9 cents. The two demand curves will differ as shown in Figure 4. It is ordinarily said that the demand for which the response of percentage change in quantity to percentage change in price is the larger (for good *A*) is *more elastic* than that of the less responsive demand (for good *B*). We will return to this idea of *elasticity* in a moment but may indicate meanwhile that the slope of a curve (on arithmetic scales) does not indicate the ratio of *percentage* quantity change to *percentage* price change and that the relation of the slope of two curves is not always an indicator of the relation of these ratios. Slope measures the ratio (inverted) of absolute changes in quantity and price, which is *per se* less significant than the corresponding ratio of percentage changes.

Demand curves may also differ in shape according to whether they are *linear* or *curved*. A *straight-line* demand curve shows that a series of successive price changes of constant amount will elicit a series of quantity changes of constant amount; this would be true, for example, if for every price reduction of 1 cent an increase in sales of 20 pounds took place. A *curved-line* demand curve shows that a series of successive price changes of constant amount will elicit a series of quantity changes of varying

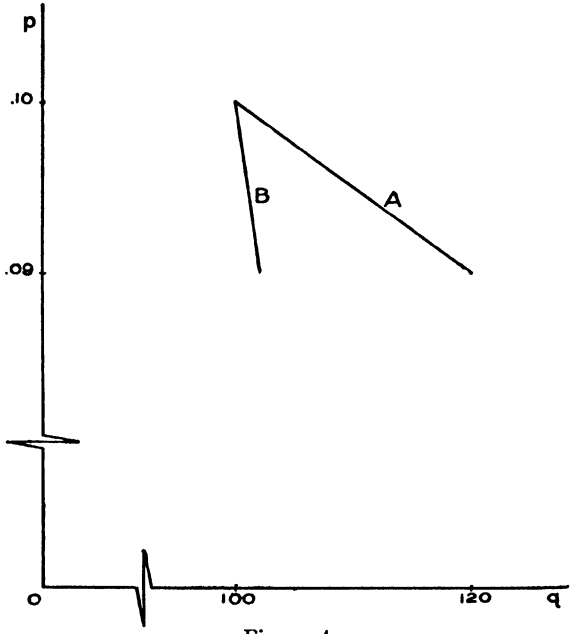

Figure 4

amount—thus five successive price reductions of 1 cent might elicit quantity increases of successively 2, 3, 4, 5, and 6 pounds. To illustrate this let us suppose the following two demand schedules for the two goods A and B:

Price of A	Quantity of sales of A	Price of B	Quantity of sales of B
10	100	10	100
9	120	9	102
8	140	8	105
7	160	7	109
6	180	6	114
5	200	5	120

If we plotted these schedules as demand curves, the demand for A would be a straight line, and that for B a curve. Demand curves will thus differ in degree of curvature as well as in steepness of slope. Some of them may also be irregular in shape.

The most significant thing about a demand curve, however, is that it shows the exact relation of each price change to the cor-

responding change in quantity bought. It also shows, inferentially, the effect of changes in price, p, on the *total revenue* from sales (p multiplied by q). Suppose we have a demand curve F which shows sales of 1000 units at the price of $1.00 and sales of 1100 at the price of $0.99; a demand curve G which shows sales of 1000 at the price of $1.00 and sales 1010.1 at the price of $0.99; and a demand curve H which shows sales of 1000 at the price of $1.00 and sales of 1005 at the price of $0.99. In addition to the fact that quantity is most responsive to price in F, less so in G, and least so in H, we may note a crucial respect in which these curves differ. *In curve* F, *total expenditure on the good increases as price is reduced.* Thus at the price of $1.00, total expenditure is $1000, but at the price of $0.99, buyers will spend $1089. A price reduction increases total expenditure, and conversely a price increase will reduce total expenditure. *In curve* G, *a price change leaves total expenditure unchanged* (to a very close approximation in this example—exactly in conception). Total expenditure is $1000 at $1.00, and also $1000 at $0.99. Finally, *in curve* H, *a price reduction results in a curtailment of total expenditure.* At $1.00, the total expenditure is $1000, but at $0.99 it is only $994.95. These curves differ generally in that price reductions result in, respectively, increase, no change, and reduction of total expenditure. The direction of this effect may be ascertained for any demand curve by multiplying together the price-quantity combination at each of a succession of prices.

The direction and rate of response of total expenditure on a good to change in its price is indicated by the *elasticity* of its demand. Where total expenditure increases with a price reduction (or decreases with a price rise)—curve F—the demand is called *elastic*. When total expenditure is constant in spite of a price change, the demand is called *unit elastic* (curve G). When total expenditure decreases with a price reduction, or increases with a price increase, the demand is called *inelastic* (curve H). This is a rough tripartite classification of elasticity. But it is obvious that among *elastic* demands, there are those which are "more elastic" than others—in some a small price reduction might elicit a rather small increase in total expenditure; in

others a small price decrease might cause a very large increase in total expenditure. Some measure not only of the *direction* of change, but of the *degree* of responsiveness of total revenue to price change is useful.

An accurate measure of the direction and degree of response of total revenue may be obtained by dividing the proportionate change in quantity of sales by the corresponding proportionate change in price for an infinitesimally small price change at any point on a demand curve. This corresponds to the percentage change in quantity divided by the corresponding percentage change in price when the changes are *very small.* Thus in Figure 5, we have a demand curve and an initial price p and quantity q. To measure the elasticity of the demand curve at the point, let us suppose a very small price reduction to p_1, by the amount of Δp. This is accompanied by an increase of quantity to q_1, by the amount of Δq. These changes are so small that the difference between p and p_1, or between q and q_1, may be neglected in calculation. Then the elasticity of demand is measured as

$$e = \frac{\dfrac{\Delta q}{q}}{\dfrac{-\Delta p}{p}} = \frac{\Delta q}{-\Delta p} \cdot \frac{p}{q},$$

where Δ stands for "a small change in."

Now when elasticity e, thus measured, is exactly -1 (the sign is necessarily negative if price and quantity change in opposite directions), total revenue will remain unchanged in response to a price change. If e is greater than -1 (-2, -3, -10, etc.), total revenue increases with a price reduction and in greater degree as e is greater. If e is smaller than -1, (-0.7, -0.5, -0.2, etc.), total revenue decreases with a price decrease and in greater degree as e is smaller. The direction of response of total revenue is thus indicated by whether e at a point is greater or less than -1. The degree of response is indicated by the size of e.

Elasticity is precisely the ratio of proportionate change in quantity to proportionate change in price at a point on a demand curve. When this ratio is taken for discrete changes in price and

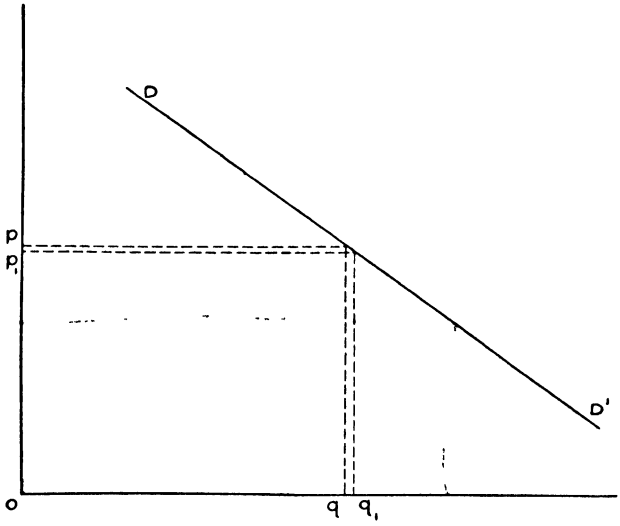

Figure 5

quantity, which are not very small, so that p and p_1 differ significantly for calculation, it does not give a precisely accurate indication of the direction and degree of response of total revenue to price change. But an approximation subject to a small error may be obtained by computing the ratio of percentage change in quantity to the corresponding percentage change in price, using in the denominators of the fractions the smaller values of both p and q. Thus in curve F above, we may get elasticity approximately as

$$e = \frac{\dfrac{\Delta q}{q}}{\dfrac{-\Delta p}{p_1}} = \frac{\dfrac{+100}{1000}}{\dfrac{-.01}{.99}} = -9.9$$

In curve G,

$$e = \frac{\dfrac{+10.1}{1000}}{\dfrac{-.01}{.99}} = \text{(approx.)} - 1.$$

In curve H

$$e = \frac{\frac{+5}{1000}}{\frac{-.01}{.99}} = -0.495.$$

The convenience either of the precise point measure or of the approximate "arc" measure is apparent. When the demand curve is shaped so that expenditure increases with price reductions, e is greater than -1 (since proportionate quantity change exceeds percentage price change). When expenditure is constant in spite of price change, e is -1 (since proportionate quantity change is equal to percentage price change). When expenditure falls with price reductions, e is less than -1 (since proportionate quantity change is less than percentage price change). By the size of e, we can tell approximately how elastic or inelastic demand is. Thus an elasticity of -50 would indicate a very elastic demand, one of -2 a moderately elastic demand, one of $-.003$ a very inelastic demand, and so forth.[8]

Two explanatory comments should be added to the preceding remarks. Since elasticity is measured by $\frac{\Delta q}{\Delta p} \cdot \frac{p}{q}$, it is clear that elasticity is related to the *slope* of the demand curve, which is measured by the inverse of $\Delta q/\Delta p$—a ratio of absolute changes in q and p. But elasticity is not simply the inverse of slope. It is this multiplied by p/q. Thus although a steeply sloped demand curve is generally evidence of inelasticity, and a gradually sloped one evidence of elasticity, the slope must be related to the relative magnitudes of p and q at the point where elasticity is measured in order to get the precise measure of elasticity.

Further, elasticity is a precise concept only as it refers to a *point* on a demand curve—*i.e.*, to an indefinitely small change in price at that point. Elasticity may obviously differ from place to place on a demand curve (and ordinarily does); each point has potentially a separate elasticity. Thus for a straight-line demand curve, such as that shown in Figure 5, we would find a different

[8] See Stigler, *op. cit.*, pp. 51-54, for a further discussion of elasticity.

elasticity at every point, as $\Delta q/\Delta p$ remained constant at all points and p/q varied, so that elasticity decreased for successively lower prices.

For purposes of further analysis in this volume, we need not labor the mathematical niceties of elasticity. We may be content to understand what it measures and why it is important. It is a precise measure of the "shape" of a demand curve at any point, and it is primarily important as an indication of the reaction of total expenditure to price change on a given demand curve.

FURTHER COMPLEXITIES OF DEMAND ANALYSIS

The preceding discussion has been concerned with the demand curve for the output of a single industry, specifically for the assumed instance of a precisely defined and neatly delimited group of sellers of identical outputs. For such an industry we have analyzed the relation of the quantity sold by the industry to the price which all firms within it charge, when there are given buyer tastes, given total purchasing power, and given prices for all other industries. It is established that for the output of such an industry under these conditions, there is a definite market demand curve, with a negative slope and with some shape which gives rise to a certain elasticity or complex of elasticities.

Each industry in turn has some such demand curve for its output, and together these industry curves constitute a family of demand curves for all output, with each curve representing, in a prevailing situation, the quantities of a specific commodity salable over a range of prices for that good. This being recognized, a first additional question is: What determines the relative positions and elasticities of the demands for the outputs of various industries?

The relative positions of various industry demand curves—that is, the absolute amounts bought over the usual range of prices and the corresponding share of all purchasing power which the industry commands—depends upon the current state of productive techniques, the state of consumer preferences, and the corresponding importance of the good in the production and consumption scheme of the economy. Thus the demand for

wheat flour may be quite large at every practically conceivable price, whereas the demand curve for fishermen's rubber hip boots may reflect by its position a very small demand at every price. The various positions of different industry demand curves is a simple matter but one of great importance in a free-enterprise economy. Any individual demand curve, reflecting buying power seeking a particular good, tends to attract business enterprises to supply that good up to some point consistent with the costs of making it and with the firms' calculations of maximum profit opportunities. Thus a very large demand for gray cotton yarn tends to elicit a large investment and the employment of a large labor force in making it; a small demand for precision barometers elicits a small investment and small employment in making barometers.

Looking at the matter more broadly, the relative or comparative sizes of the various demands for all goods (together with the relative costs of producing them) tend to apportion or allocate productive effort among various goods roughly according to the relative spending power buyers are prepared to offer for them. Business enterprises are not permanently wedded to the production of certain goods; over long periods they shift willingly from one product to another according to the money demand for it and the profit opportunities it offers. New enterprises will be brought into being, moreover, to supply new or developing demands. The complex family of individual demands for goods tends, therefore, to guide the apportionment or allocation of resources into various lines of production. We must consequently keep it in mind that any seller, though immediately faced by the market demand for the good he is currently producing, is in his long-range planning faced by a family of demands for all goods, the relative attractiveness of which he must keep in mind in considering his profit opportunities. This observation will appear to be of fundamental importance when we later consider the function of the economy as a whole.

The shape or elasticity of an industry demand curve in general reflects the substitution relationships between the output of the industry in question and those of other industries. When any such curve is drawn on the assumption of all other prices being constant, it essentially reflects the extent to which the

immediate good is substituted for others as its price is reduced, or to which others are substituted for it as its price is increased. It is apparent that if a good has no adequate substitutes at all—if it furnishes a satisfaction which cannot be obtained alternatively from other goods or replaced by other satisfactions—it will tend to have an inelastic demand. Thus tobacco products are ordinarily reckoned to have an inelastic demand. Increases in their prices up to a rather high level will not reduce their sales by much, since smokers have no other product to turn to. Correspondingly, price reductions on tobacco products will not stimulate their sales very much. On the other hand, goods with a number of substitutes may have much more elastic demands.

The elasticity of the demand curve of an industry, of course, depends in part on whether or not other industry prices respond significantly to changes in its price. The ordinary assumption is that no other price responds enough to influence the industry's demand—that is, the good has no very close substitute but is moderately substitutable for each of quite a range of products. If this is the case, there are definite limitations on the elasticity of an industry demand curve. There are cases, however, of interrelated or quasi-separate, industries—such as sellers of theater admissions and sellers of night-club entertainment—the demand for each of which is significantly influenced by the price of the other. In this event, neither demand curve is validly constructed on the assumption of the other price being constant but should instead be construed as showing the response of sales to price in one industry given any systematic covariation in the other price. Such price responses by a substitute output of course reduce the elasticity of demand for an industry. More generally, of course, we should recognize that the two industries are interdependent and should take explicit account of the interdependent character of their respective demands. In a great many cases, however, we find industries more clearly defined, with the price of the immediate industry having a negligible effect on the demand for any one outside industry. It is therefore ordinarily valid to assume that no other price will respond significantly to change in the immediate industry price and to construct the industry demand curve on this assumption.

The close interdependence of industries is not limited to cases of substitutability, since there are also cases of *complementarity* of goods—instances where a decline in the price of one good, by stimulating its use, will also stimulate the use of another good ordinarily used together with it. Hardwood flooring and plumbing fixtures—both building materials—are complementary goods. A fall in the price of flooring, so far as it stimulates building by reducing its cost, also stimulates the demand for plumbing fixtures—shifts the demand curve for them to the right. Such groupings of interrelated complementary demands are fairly common. Where the complementary effect is large enough to set up perceptible price repercussions between industries, either demand curve must be drawn given the systematic covariations in the other price, and the interdependence of the prices should be recognized.

Subject to the exceptions of industries closely related by the perceptible substitutability or complementarity of their outputs, we have an economy made up of a large number of relatively independent industries. The demand curve for any one of them may be drawn on the assumption that the prices of other goods are given and unchanging, or, more precisely, on the assumption that no other price will change enough in response to a change in this industry's price to influence the sales of this industry perceptibly. This is in part because substitutability and complementarity between industries are in general not too close, and in part because most industries secure a relatively small proportion of the total purchasing power expended on all goods. Although the interrelationship of the prices and outputs of most industries may thus be insignificant or imperceptible from the standpoint of sellers in any one of them, it is nevertheless true that all industry demands are in some degree interdependent. The position of the demand curve for any one good depends directly upon the prices of all other goods, although it may in many cases depend insignificantly on the price of any one other good. No single industry demand is genuinely independent. Looking at the economy as a whole, we have an essentially interdependent family of industry demand curves, the positions of which are mutually determined and interdependent. The firms of industries pursue their price adjustments in disregard of this inter-

dependence, since their individual impacts on the total situation are negligible. But from the concert of independent industry price adjustments are determined the relative positions of all the industry demand curves in the economy.

It must of course be recognized that the position of every demand curve depends upon the size of the aggregate flow of purchasing power seeking all goods. For most goods, a larger general purchasing power means a larger demand. Most demand curves shift to the right as money incomes expand during the upswing of the business cycle, and to the left with the decline in general purchasing power which characterizes business recessions.[9] The over-all volume of purchasing power is thus a strategic determinant of the incentive for enterprisers to produce in any line.

Not all demands, of course, respond to shifts in general purchasing power in the same proportion. The demands for luxuries and for durable goods increase most rapidly with increases of general income and fall off most heavily as income declines. The pattern of relative demands for all goods thus shifts systematically with shifts in the general level of income.

So far in the immediately preceding discussion, we have proceeded on the simplifying assumption that within each industry the member firms sell identical outputs at a necessarily identical price. This has allowed us to speak of the industry as having at any one time a single commodity and a single price common to all sellers, and thus has enabled us to construct a simple industry demand curve. The argument is made more complex, although its fundamental conclusions are not significantly altered, when we recognize that within many industries the outputs of various member firms are not identical but differentiated and that their prices are thus not necessarily identical. Thus in the soap industry, one seller may price his product at 40 cents a pound, another at 38 cents, another at 42 cents, and because of brand preferences all may succeed in selling soap at the same

[9] There are certain exceptional ("inferior") goods the demand for which may decline as income increases. This is said to be true of bread in countries of low income. Poverty forces people to subsist on little else when income is low, but other foods may be substituted for it rapidly as income increases.

time. Where such product differentiation occurs, it is not possible in strict logic to construct an industry demand curve relating *the* price of the industry to the total output. There may be more than one price, and the outputs do not add to a strictly homogeneous total. It is nevertheless possible to construct a provisional or working demand curve for the industry which represents the change in the aggregate amount (neglecting nonhomogeneity) which all firms can sell if, starting with any given set of interfirm price differentials, they concurrently change their prices by identical proportions. Some such approximation must stand in lieu of a more satisfactory industry demand curve in such cases. It is in fact a relatively satisfactory substitute, since the product differentiation within industries is usually not so great as to make summation of different outputs unreasonable, and since in most cases the member firms will be held by competition to very similar if not identical prices. The industry demand curve in such industries is thus a slightly arbitrary concept, but it is nevertheless quite valid and in no wise upsets the generalizations heretofore drawn concerning the relation between different industry demands or the close interdependence of firms within an industry. Further implications of product differentiation will be discussed as we turn to the demand curve for the output of the individual firm.

THE DEMANDS FOR INDIVIDUAL SELLERS' OUTPUTS

The industries the demand curves for which have been discussed above are, of course, nothing more than groups of firms whose prices are closely related because they sell close substitute outputs to a common group of buyers. Such industry demand curves show how much all buyers from the industry will take at each of a range of alternative prices during a given time interval. As such, it is the demand which all the firms in the industry face together—the common demand for their combined output. If there are 75 sellers in an industry, and the industry demand curve shows sales of 10,000 at the price of $1.00 and of 11,000 at $0.95, this means that if the combined offerings of all 75 sellers are 10,000 units, the price for their

output will be $1.00 per unit, but if their combined offerings are 11,000, the price will fall to $0.95.

If an individual seller of a good knows the industry demand curve for this good—or makes an estimate of it on which he depends—it gives him the answer to one question: If (under given circumstances of buyer income, etc.) the total supply of the product is any amount q, what will the market price be? If the seller now wants to estimate future market supply, he can in turn estimate the price which he and his competitors will get for their outputs. Knowledge of market demand curves may thus be useful to individual sellers, and in some types of industry individual sellers may take direct account of the industry demand in making their decisions regarding price and output.

The industry demand curve for a commodity, however, does not necessarily allow the individual seller to answer a second question: What will his own selling price be if his output is alternatively the amounts x_1, x_2, x_3? In short, it does not necessarily relate the individual seller's output to the price he will receive. Yet every seller should be primarily interested in a curve which will give him just this information. He wants to know his *individual seller's demand curve*, which may be defined as a curve showing how much output the individual firm can sell at each possible price.

The demand for a single firm's output is evidently derived from or related to the industry demand for the sort of product the firm produces, and is determined by the character of that industry demand and by the competitive relation between the instant firm and other firms producing identical or close substitute products. With some governing industry demand, the relation of price charged to quantity sold by a single seller depends primarily on the character of competition in his industry. Common-sense observation tells us, moreover, that there are a number of distinctly different sorts of competitive situations and that the character of the seller's demand (and the way it is related to governing industry demand) may thus differ considerably from industry to industry. In fact, the firm's calculation of its own demand may be of several distinct types, and a first step in considering how firms calculate the demands for their

products (and eventually their price and output policies) is to construct *a classification of industries.*

Industry or market classifications may be simple or complex, depending upon the number of differences among industries which we recognize. For beginning purposes we may distinguish five classes, as follows[10]:

1. Industries with one seller.
2. Industries with many sellers.
   a. Where the products of all sellers are identical.
   b. Where the products of various sellers are "differentiated."
3. Industries with a few sellers.
   a. Where the products of all sellers are identical.
   b. Where the products of various sellers are "differentiated."

It will be noted that in this classification markets are distinguished according to only two characteristics: the number of sellers—many, few, or one; and the relation among the outputs of the sellers in an industry—whether they are identical or differentiated. The meaning of a "one-seller" industry is evident when we have defined an industry. By "many" sellers is meant a large enough number that each is so small that he controls an insignificant proportion of the total industry output or little enough that any possible change in his output will add or subtract so little from the total industry output as not perceptibly to affect the industry price. By "few" sellers is meant a small enough number that one or more sellers controls a significant proportion of industry output, so that changes in his output may add or subtract enough from industry output perceptibly to affect the industry price. The simplest version of fewness of sellers is found where there is a small number of firms (for example two, ten, or twenty) of approximately equal size, so that each controls a significant share of industry output. The meaning of product differentiation has already been discussed. In some industries, the outputs of competitive sellers will be

[10] Cf. Fritz Machlup, "Monopoly and Competition," *American Economic Review*, September 1937, pp. 445-451.

substantially identical, so that buyers make no distinction or choice among the product offered by different sellers. In others, the outputs of the various firms, although close substitutes to buyers, are not perfect substitutes, being distinguished by design, quality, packaging, "firm name" or branding, or sales promotion.

It will appear that these two characteristics of industries are probably the most important in determining the relation of an individual firm's demand to that of its industry (and thus to other demands in the economy). The number of sellers is significant because it determines the extent to which individual sellers take direct account of the industry demand and of the possible reactions of their competitors in determining their own prices and outputs. The degree of product differentiation is important because it determines to what extent the firm enjoys some potential independence in pricing its product, and some means of directly influencing or maintaining its own volume of sales. There are, of course, other characteristics of industry with a significant influence on the individual firm's demand—among them the degree of concentration of output among sellers (proportions of industry output controlled by the several sellers), the degree of durability in use of the output, and the number of buyers.[11] For the moment, however, we shall concentrate attention on the significance of the number of sellers and the degree of product differentiation.

SINGLE-FIRM MONOPOLY

The firm's view of the relation of its sales volume to the price it charges is perhaps simplest in an industry with one seller—in so-called "single-firm monopoly." A single-firm monopoly is most easily defined as the case where one firm has an output for which there is no one substitute the price of which will be influenced enough by the monopolist's price changes to cause the monopolist's demand to shift perceptibly, and for which there is also no related range of close substitutes concurrent changes in

[11] See E. S. Mason, "Price and Production Policies of Large-Scale Enterprise," *American Economic Review*, Supplement, March 1939, pp. 61-74.

the price of which greatly affect the monopolist's demand. The single-firm monopolist has no direct rival and also faces no group of closely competitive outputs. The single-firm monopoly category is relatively rare in our economy, but several industries, including cash registers, basic aluminum, and telephone communications, are or once were practically monopolized.

The relation of the sales volume to the price of a single-firm monopolist is evidently shown by an industry demand curve for his product—his own "seller's" demand curve is an industry demand curve for an entire commodity. The characteristics of demands for the outputs of such monopolists are therefore those of industry demand curves generally, as regards both direction of slope and elasticity; and they may differ from monopoly to monopoly according to the varying patterns of buyers' desires for the various products. The demand for a single-firm monopolist's output of a certain product might appear as in Figure 5.

The fact that a single-firm monopolist, in calculating the effect of each possible price change on his sales volume, can refer to the industry demand for an entire product is important in two principal ways. First, the monopolist faces a demand schedule for his output which, subject only to his finding it out, is quite definite. He can freely select each of a large number of prices, with the expectation that a definite sales volume will correspond to each. He can move from one price to another without directly engendering any competitive reactions in other prices which which would in turn influence his sales. This is because, as a single-firm monopolist, he produces a good for which there are no "close substitutes"—*i.e.*, no substitutes sufficiently close that a price change by the monopolist will elicit price changes in direct response which will perceptibly influence the monopolist's demand. And his demand will not shift about greatly in response to the independent price changes of any limited range of close substitutes.[12]

[12] Is the definition of single-firm monopoly thus adopted too narrow? Evidently we must exclude from the category firms for whose outputs there are many close substitutes, concurrent movements in the prices of which will cause the firm's demand curve to shift markedly and systematically. (This situation is better described in a special category as monopolistic competition.) But we have also excluded cases where a single seller has one or a few substitutes for

The single-firm monopolist, to summarize our first point, has a definite and relatively stable demand schedule for his own output. The second significant property of this monopolist's demand schedule is that, like any industry demand curve, it sloped downward to the right. This means, from the monopolist's standpoint, that under given circumstances (of general purchasing power and the like) an increase in the output he offers to buyers will result in a reduction of price; a decrease in output should cause an increase in price. The monopolist's revenue per unit of output tends to fall as output increases. As a result the monopolist, even if his output cost him nothing, would have some tendency to restrict his production below the maximum, seeking that combination of price and output which promised the greatest *aggregate* profit. In short, the monopolist has a definite demand curve for his output, with some degree of slope downward to the right; he is therefore in a position to exploit the given relation of price to output to his maximum advantage. In effect, he can deliberately select a *price policy* of his own.

his output sufficiently close that their prices shift in response to changes in his price, and by shifting perceptibly influence the volume of his sales. Such a situation is excluded from single-firm monopoly generally because, where such direct price interdependence exists, it is ordinarily mutually recognized by the sellers involved, a direct rivalry exists, and no seller can know his own demand curve without guessing his rivals' reactions to his own price changes. Since such anticipations, especially when the correctness of *A*'s anticipations of *B*'s actions implicitly depend on the character of *B*'s anticipations of *A*'s actions, are often extremely indefinite and uncertain, the individual seller then does not have a definite or unique demand curve for his output, and the situation is better described as oligopoly than as monopoly. There may be limiting cases, however, where a single seller, acting independently, may legitimately anticipate unique, definite, and limited price reactions on the part of several substitute outputs when he changes his own price (in turn affecting his own demand in determinate fashion), so that he can predict definitely the total effect of his own price change on his own output. This might be true if the prices of several substitutes always reacted in a definite and more-or-less automatic fashion to the monopolist's prices, without significant uncertainty appearing. Such limiting cases may also be referred to as single-firm monopoly, the monopolist still retaining a unique and determinate demand curve for his own output. We will view single-firm monopoly in general, however, as typically the case of a seller with an output having no close substitutes and no perceptible price interdependence with any other one output. Cf. Stigler, *op. cit.*, pp. 219-221.

The manner and the extent to which the monopolist exploits the demand for his product will of course depend upon its elasticity. If he has a very elastic demand, output may be extended with very small sacrifice in price, and there will be less tendency to restrict output. With a very inelastic demand, on the other hand, where total revenue is greater the smaller the output, there may be a tendency to restrict output indefinitely or until the demand finally becomes elastic.

The preceding is a partial analysis of the demand for a single-firm monopolist's product. To understand its full significance, we must consider how such a monopolist balances the cost of his output against this demand to select a definite price and output which will maximize his net profit. Before turning to this (see Chapter 5), however, we should consider the relation of price to output as it is viewed by firms in other categories of markets.

PURE COMPETITION

Monopoly, in the sense of full-fledged single-firm monopoly which can disregard the rivalrous reactions of other sellers, is rare in the American economy. This is also true of the markets at the opposite extreme, with many sellers, all of whom sell a homogeneous (identical) product, often characterized as markets in *pure competition*. An industry in pure competition has enough sellers, all relatively small, that no one of them produces a significant proportion of the total market supply. No seller, either by extending his own output to the practical limit or by withdrawing it entirely, can perceptibly influence the market price of his product. The product of every seller, moreover, is in the minds of buyers quite indistinguishable from that of every other. Ordinarily a hundred or more small sellers, none of whom produced more than 1 or 2 percent of the market output, would be required to fulfill approximately the condition regarding numbers; and all would have to produce some single commodity sold at specified grades and quite undifferentiated by quality, advertising, packaging, or branding. We find a fair approximation to pure competition in agricultural markets for certain grain crops, and in the industrial field in the cotton gray goods industry. But the practical instances of pure competition

are relatively few. It may nevertheless be instructive to observe the contrast between the single seller's demand curve in pure competition and in monopoly.

The seller in pure competition, since he produces a very small and indistinguishable proportion of a large aggregate supply, has no control over the price of his output. The going market price for the good he makes (*e.g.*, gray cotton yarn) is the price at which he can sell his maximum output or any part of it. He cannot force market price up perceptibly by withholding his supply or depress price by increasing his output within practical limits. To look at it in another way, he can sell nothing at any price above the market price; he can sell all he can produce at the market price; therefore he has no reason to sell below the market price. If we diagram the resulting relationship of his output to his selling price, we find that *in pure competition the demand curve for a single seller's output is a horizontal straight line at the level of market price*, as in Figure 6. In effect, as the seller views demand, changes in his output have no influence on price; it is not possible for him to initiate price changes. He can set his output at any attainable level without influencing his selling price. The individual seller's demand curve is related to the industry demand so that at any time it lies at that level of price determined by the interaction of the industry demand curve and the aggregate market supply offered by the many firms of the industry.

The demand situation faced by the small seller in pure competition thus differs from that faced by the monopolist in two significant ways. First, the competitive seller has no control over price and no possibility of adopting a price policy. He can at the most select the output he wants to produce at the going price. Since increases in his own output do not tend perceptibly to reduce his price, he does not have the virtual tendency that the monopolist does to restrict his output.[13] Only the rise in his costs with increases in his output will limit his production. In short he is not in any position to exploit for his own ends the slope

[13] The horizontal demand curve is "perfectly elastic," or has elasticity equal to infinity, since there is indefinite increase in quantity with zero decrease in price.

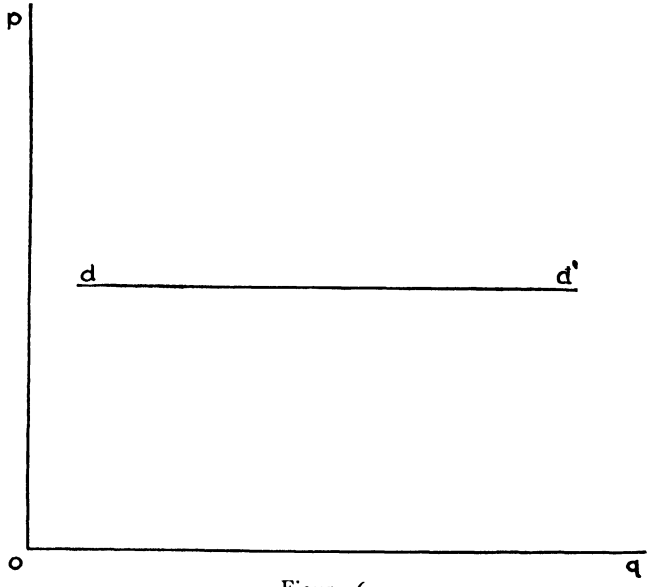

Figure 6

of the industry demand curve for the good he produces. A second peculiarity of demand situation of the seller in pure competition is that his own demand curve is much less stable than that of the monopolist. Since it is just a line at the level of the going market price, it shifts with every change in market price and thus in response to any shift either in the industry demand curve or in the volume of market supply. By contrast, the monopolist's demand curve changes only with shifts in the industry demand schedule (resulting from changes in buyers' income, tastes, or the prices of other goods), but is obviously not affected by competitive changes in the supply of the good he produces. The seller in pure competition thus has a less stable and dependable demand curve on which to base his calculations. The preceding is sometimes aptly expressed by saying that the seller in pure competition is simply at the mercy of an impersonal market price, which changes frequently and is outside his power to control, whereas the monopolist is able to "administer" price, or to select one of many alternative prices according to his own best interests.

MONOPOLISTIC COMPETITION

Neither single-firm monopoly nor pure competition is especially common in the American economy; these categories represent mainly extreme cases between the limits of which most real markets lie. In practice the number of sellers in an industry is usually neither *one* nor *very many*, and the products they sell are usually not identical, but differentiated in some degree. Thus we find that the three remaining categories of our market classification are factually more important than those already mentioned. These three categories are:

1. Industries with few sellers selling identical products, or *pure oligopoly*.
2. Industries with few sellers selling differentiated products, or *differentiated oligopoly*.
3. Industries with many sellers and differentiated products, or *monopolistic competition*.[14]

Of these the second is probably the most common and the third the least common, but there are numerous industries which have the approximate characteristics of each category. Let us investigate the character of a seller's demand curve in each case.

An industry in monopolistic competition has many sellers with differentiated but close-substitute products. By "many sellers" we mean that the sellers are many and small enough that no one of them controls a significant proportion of the total market output; no firm by extending or reducing its output will affect the sales of any other seller enough to induce a direct reaction. (This would be true, for example, if there were 100 small sellers. Then the aggregate effect on other sellers of a gain in sales by any one would tend to be divided 99 ways, with no one effect—or even the total effect—being noticeable.)

The products of the various sellers are relatively *close substitutes* for each other (as one brand of cigarettes for another) but they are not *perfect substitutes*. Monopolistic competition is thus

[14] "Monopolistic competition" is used here to refer to a somewhat narrow and restricted market category. It may also be used to refer to all pricing under other than purely competitive conditions. See Chamberlin, *op. cit.*

distinguished from pure competition, where the products of all the sellers in the group are perfect substitutes. Since an individual seller in monopolistic competition has a product which no other seller duplicates, at any rate exactly, he is in a sense a *monopolist*. The practical difference between his position and that of the single-firm monopolist described above is that the seller in monopolistic competition has a product for which there are many close substitutes, whereas the product of the single-firm monopolist has no such substitutes. Consequently, the demand for the monopolistic competitor's product will be much more sensitive to a relatively small range of prices than that of the single-firm monopolist. His position is like that of the typical single-firm monopolist in that his price does not influence the price of any one other seller enough to engender a significant reaction, and he thus disregards other prices in setting his own. It is unlike the monopolist's in that there is a closely related group of other sellers concurrent changes in whose prices will definitely influence his demand curve—thus the demand for his output is much less independent and stable than that of the monopolist. The monopolistic competitor stands in much the same relation to an industry of close-substitute outputs as that of the single-firm monopolist to the economy as a whole. We recognize the industry in monopolistic competition as a separate case because in practice we frequently find a group of closely related sellers producing close-substitute though not identical products, and such a group is separated from all other sellers by a rather wide gap in intersubstitutability of products. Such a group of closely related sellers, if it is large, is designated an "industry" in monopolistic competition.

The relationship of price to output for a seller in monopolistic competition thus reflects the fact that he is a sort of a monopolist, or that he possesses a "degree" of monopoly in his distinctive product. At least some buyers distinguish his product from close-substitute products, rating it as slightly preferable or slightly less desirable. As a result a small reduction in this seller's price (the price of close substitutes remaining the same—and they will not respond to one seller's price change) should increase his sales volume substantially, and correspondingly a slight increase in price should reduce his sales greatly. In effect

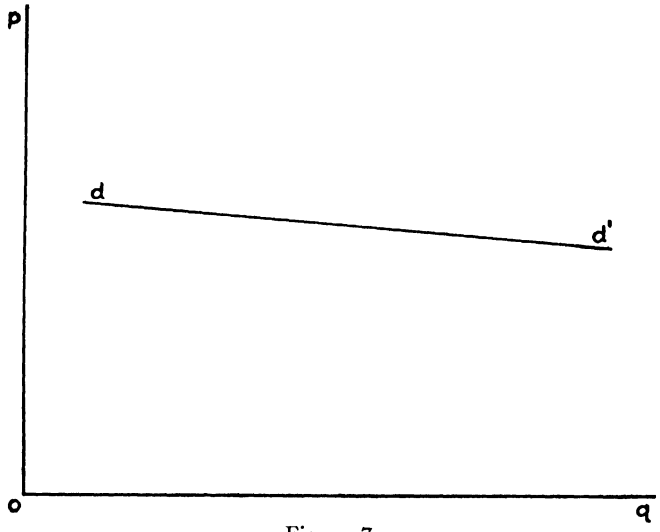

Figure 7

the demand curve for the individual seller in monopolistic competition should ordinarily be one which slopes downward to the right very gradually, or is *very elastic* to price changes, like that shown in Figure 7. This is in contrast to the demand curve of the single-firm monopolist, which should ordinarily be less elastic. With so very elastic a demand curve for his product, the seller in monopolistic competition, although he can exercise some choice over his price, in fact finds this choice limited to a very small range of prices.

The second distinction between the demand curve for a seller in monopolistic competition and that for a single-firm monopolist is that the position of the former is much more sensitive to a limited number of other prices than the latter. The industry demand curve for any product of course tends to shift in response to changes in the price of substitutes—outward if substitutes become more expensive, backward if they become cheaper. The position of the demand curve for every good in effect depends upon the pattern of prices of all other goods. This is as true of a single-firm monopolist's demand curve as it is of that of a monopolistic competitor. But the single-firm monopolist

typically sells a product for which there are only a very large number of relatively distant substitutes, and his demand curve shifts only mildly in response to changes in the prices of any one or few of them. In monopolistic competition, on the other hand, the seller's demand curve shifts immediately and clearly in response to concurrent changes in the prices of a limited group of close substitutes. It is therefore much less stable—much more a will-o'-the-wisp. This is another way of saying that in monopolistic competition the seller comes very close to being at the mercy of a market price for the group of which he is a member, and enjoys only a slight degree of independence in pricing.[15] The position of his demand curve is determined by the prices and outputs of all sellers in his industry; the relationship of the aggregate output of all sellers of the group to the *industry* demand determines the general position, on the price scale, of the demand curve for the output of each seller.

OLIGOPOLY

The three market situations discussed so far, although they evidence significant differences, have one thing in common. In each case, the individual firm has a "definite" demand curve for its product; it can postulate and potentially learn a definite relationship between the price it can charge and the output it offers, and it can decide to produce a certain output with the assurance that its own particular decision will not induce any *retaliatory* changes in the prices of substitute products. In the case of single-firm monopoly, this is typically so because there is no substitute product "close enough" to have its sales significantly affected by the monopolist's price changes. In pure or monopolistic competition it is because the competing sellers are so numerous and small that no one of them commands a significant proportion of the total market of the industry and thus cannot influence the sales of any one rival perceptibly. But in any event, no seller in any of these three categories has a demand curve which will shift about more or less unpredictably in response to competitive price *retaliations* if he changes his own

[15] See Chamberlin, *op. cit.*, Chap. 5.

price or output. There is no *mutually recognized interdependence* of individual sellers' price or output policies.

The situation is definitely otherwise in those dominant market categories where the sellers are *few* in number—"pure oligopoly" and "differentiated oligopoly." These two groups embrace all those industries where the competing firms are several but are few enough that each controls enough of the total market for the sort of good they produce that an extension of its output (or a reduction in its price) will tend to reduce the sales of rival firms by a noticeable amount. This is definitely the case if from two or three up to perhaps two dozen firms control the bulk of an industry, and it holds true to a degree until the number of firms becomes quite large. In a large proportion of all of our industrial markets there is "fewness of sellers" in this sense. Either the entire output of an industry is controlled by six, seven, ten, or a dozen firms, or the bulk of the industry's output (say from 60 to 90 percent) is controlled by a very few large firms, while the remaining output is spread among 20 or 30 "small fry." In either event, the market structure is dominantly oligopolistic, in that the leading pricing decisions are made by a few large firms the individual demands for whose products are closely interdependent.

The dominance of this sort of market structure is, of course, a reflection of the high degree of concentration in American industry, which in turn stems from the widespread organization of business firms for the exploitation of mass-production and mass-distribution techniques. Among the "pure oligopolies" (concentrated industries producing practically homogeneous products) we would count the steel, copper, aluminum, lead, cement, industrial alcohol, sulfur, and many other industries. Oligopolies producing differentiated products include the automobile, rubber tire, electric appliance, cigarette, and soap industries. Oligopolistic industries account for much more of our industrial output than all other types combined.[16]

The dominant feature of the relationship of the price to the sales volume of an individual seller in oligopoly is its recog-

[16] See National Resources Committee, *Structure of the American Economy*, Part 1, 1939, pp. 239-263, for data on number of sellers and concentration.

nized direct interdependence to a significant degree with that of rival sellers. Of the total market demand for the sort of product an oligopolist produces, he can depend upon selling no definite independent proportion. The several sellers in an oligopoly at any moment will share total demand for the good in some given proportion which depends upon the relationships of their various prices and upon other circumstances influencing purchasers. But should any one attempt to increase his sales by an independent price reduction, the result will be uncertain. As his sales increase, other sellers will feel a definite pinch and will tend to react in some way. If they reduce price, this will reduce the demand of the first seller (cause it to shift backward), and this may in turn cause him to react again. Any independent price change by an oligopolist tends to set off a long chain of repercussions, ordinarily with no definitely predictable outcome, since the adjustments of each seller depend, in turn, upon his conjectures regarding the counterreactions of each of his rivals. In short, the oligopolist if acting quite independently ordinarily does not have a definite demand curve for his own output. He has at his going price a share of a total market demand which depends upon the price and competitive balance struck between him and his rivals. At possible alternative prices, the share he will get depends upon the course his rivals take in reaction to any changes which he makes. An oligopoly is a recognizedly interdependent group of sellers.

The dominant motif in oligopoly is therefore uncertainty. As long as the sellers in an oligopoly remain independent of each other and have no explicit or tacit means of coordinating their pricing policies, the individual oligopolist has no definite or unique demand curve to relate his price to his sales. In place of a definite demand curve, of course, the independent oligopolist may contemplate various *provisional* demand curves drawn according to various assumptions about his rivals' policies. There is logically an indefinitely large number of possible provisional or conjectural demand curves which an oligopolist might imagine.[17]

[17] See Chamberlin, *op. cit.*, Chap. 3; also R. F. Kahn, "The Problem of Duopoly," *Economic Journal*, March 1937.

For practical purposes, however, it should serve to emphasize two or three principal possibilities.

In pure oligopoly, where all sellers have identical products, they must in general charge identical prices to remain in business. The main uncertainty on the part of a single seller regarding the effect of his independent price reductions, therefore, will concern whether they will just be matched by his rivals or whether his price cuts will set off chains of retaliatory price cuts which he must match. If the latter will occur, the idea of a seller's demand curve simply gives way to a calculation of the net outcome of a price war. Uncertainty may also be felt with respect to the effect of independent price increases. If, on the other hand, price changes will just be matched, the individual seller, assuming concurrent actions by his rivals, may conceive of his own demand curve as some conventional *share* of the industry demand curve.

Because oligopolistic uncertainty may stand in the way of profitable price adjustments, there is, in fact, a tendency in pure oligopolies for rival sellers to make agreements or follow pricing conventions designed to coordinate their pricing policies and to secure concurrent action on all price changes. In the United States, where specific market-sharing and price-fixing agreements are illegal under the antitrust laws, the most usual practice is for the several sellers in an oligopoly to recognize a *price leader*, who assumes the initiative in raising or lowering price, and whose lead is promptly followed by the other firms. In this case the oligopoly becomes a quasi-unified group of sellers who view the *joint* (industry) demand for their several outputs as a *shared* demand curve, and thus becomes similar to a single-firm monopoly. Some of the main possibilities for the price-sales calculations of a single seller in pure oligopoly are therefore: (1) independent calculation of the consequences of initiating a price war; (2) independent (but uncertain) assumption that all rivals will match prices and tacitly share the market—in effect, assumption of a share of the industry demand curve as a seller's demand curve; (3) collusive adoption, in concert with "rivals," of the over-all market demand curve for joint exploitation "as in monopoly"; (4) simply uncertainty about effect of changing price. It should be recognized, moreover, that in a given oli-

gopoly the sellers may shift in unstable fashion from one sort of calculation to another over time.

In differentiated oligopoly, where a few sellers have differentiated products, the general demand situation faced by any seller is much the same, and the various specific calculations he may make follow about the same pattern. The main differences from pure oligopoly are that if the product differentiation is distinct the several sellers may find it feasible to charge somewhat different prices, and that their relative shares of the market at any price will tend to be more stable. One seller may be able to make small independent price changes without eliciting automatic reactions from rivals, and to this extent he can have an independent price policy within some narrow range. In the main, however, the alternative demand calculations possible for a seller in differentiated oligopoly are roughly similar to those found in pure oligopoly. The principal distinctions between the two sorts of oligopoly are found in the differing importance of selling costs and will be discussed later.

The economy as a whole includes industries of all types, so that in practice there are firms subject to every type of intra-group relationship and of relationship to industry demand. There are industries in monopolistic competition, pure competition, single-firm monopoly, and both sorts of oligopoly. Further, there may be firms in in-between or mixed situations. We have already emphasized that the various industry demand curves in the economy have mutually interdependent positions, so that the price-quantity of sales relation for each depends on the prices of all others. Within industries, individual firms have various types of individual demand curves for their outputs, evidencing various relations to each other and to the over-all industry demand. More broadly, the individual demand curves of all sellers in all industries constitute a mutually interdependent family, and their positions are mutually determined. For individual firms, however, the primary interdependence is within the industry. As between firms in different industries, the interdependence is indirect and largely via the behavior of the industry or group prices in question. The complex of interrelated industry and firm demands is the primary guiding force in the allocation of resources in a free-enterprise economy.

SUMMARY

At the outset of this volume we posed the question of how prices and outputs are determined by profit-seeking firms in a capitalist economy. This led us to inquire first into the individual seller's view of the relation of the price he charges to his sales volume.

Investigating this matter, we have seen first that for every sort of product there is at any time an *industry demand curve* which represents a certain relation between the price of the good and the amount of it that buyers will purchase. Although each such demand curve is generally interdependent with all others, each may be viewed as provisionally independent as long as there is no close and recognized interdependence with any one other.

The relation of the individual seller's price to his sales ordinarily is derived or stems from the industry demand for the sort of product he produces. The relationship which the seller conceives between his sales and his price (represented in the seller's demand curve) depends primarily on the type of market in which he sells. Industries differ from one another most significantly in number of sellers and in degree of differentiation among rival sellers' products. Correspondingly we have recognized provisionally at least five distinct sorts of market structure: single-firm monopoly, pure competition, monopolistic competition, pure oligopoly, differentiated oligopoly. Each market structure has a unique significance because it is reflected in certain peculiar demand conditions for the individual seller. The seller's calculation of how his sales volume is related to the price he charges is different in each of these types of markets.

In single-firm monopoly, the seller envisages this price-sales relationship as a relatively stable and independent market demand curve, usually far from perfectly elastic. He can therefore have a clear-cut independent price policy, or deliberately select a certain price-output combination. In pure competition, the individual seller sees no variation of price in response to changes in his output and considers himself able to sell all he wants to produce (in view of cost considerations) at the going

market price. His "demand curve" is thus a horizontal straight line at the level of market price, albeit one which tends to move up and down frequently. He has no "price policy" but merely selects an output in the light of market price and cost.

In monopolistic competition, the individual seller is also rather fully at the mercy of a given general level of price for the sort of product he sells and cannot successfully get very far away from this level. But at any such level he has, within a narrow range, some power of choice over his own price because of the distinctive characters of his product and of competing products. We summarize this situation by saying he has a sloping but very elastic demand curve for his own product, the position of which (regarding general level of price) is closely interdependent with the positions of all rival sellers' demand curves.

In pure oligopoly, the individual seller has no definite demand curve of his own; there is no certain fashion in which he can independently calculate the relation of the price he charges to the volume of his sales. He therefore either sets his prices on the basis of uncertain conjectures concerning his rivals' reactions to his own decisions, or, because this is an undesirable alternative to most businessmen, he arrives at agreements or tacit understandings with his rivals to cooperate in exploiting the total market demand. In differentiated oligopolies, the seller's calculations are much the same although he can allow himself somewhat more leeway in independent pricing decisions. He still has no determinate independent demand curve, unless one based on potentially unstable understandings with his rivals that they will make concurrent price decisions.

Because the seller's calculation of his price-sales relationship differs so distinctly among market categories, we are led to expect that there may be correspondingly distinct differences in pricing. It follows that our investigation of how prices are made in modern business will in effect be several inquiries into how they are made in several different industry situations.[18]

[18] Some writers, bothered by the lack of logical nicety involved in the use of the "industry" concept in a world where industries tend to overlap or have indefinite boundaries, would place less initial emphasis in analyzing demand upon industry groupings of firms. Instead they would emphasize principally the

Before we can enter fully into these inquiries, however, we must consider a second sort of determining calculation which sellers make—that of the relation of their outputs to their costs of production.

SUPPLEMENTARY READINGS

ALFRED MARSHALL, *Principles of Economics* (8th ed.), Book III.

EDWARD H. CHAMBERLIN, *The Theory of Monopolistic Competition* (5th ed.), Cambridge, Mass.: Harvard University Press, 1946, Chaps. 3-4.

J. R. HICKS, *Value and Capital*, Oxford University Press, London, 1939, Part I.

HENRY SCHULTZ, *The Theory and Measurement of Demand*, Chicago: University of Chicago Press, 1938.

ROBERT TRIFFIN, *Monopolistic Competition and General Equilibrium Theory*, Cambridge, Mass.: Harvard University Press, 1940.

GEORGE J. STIGLER, *The Theory of Price*, New York: The Macmillan Company, 1946, Chaps. 4-6.

RUBY TURNER NORRIS, *The Theory of Consumer's Demand*, New Haven, Conn.: Yale University Press, 1941.

demand for the output of the individual firm and the relation of each firm's demand to the prices of other firms' outputs (measured by the so-called cross-elasticity of demand). This approach, by attempting less than the "industry" approach, avoids the use of imprecisely defined concepts. As such a demand analysis is employed in analyzing price determination, however, it must eventually refer to closely interdependent groups of firms, or provisional "industries." Careful and critical employment of either approach should lead to satisfactory and similar results. For a discussion of the alternative approach mentioned, see Robert Triffin, *Monopolistic Competition and General Equilibrium Theory* (Cambridge, Mass., 1940).

CHAPTER 3

THE PRODUCTION COSTS OF THE FIRM

The decision of the individual enterprise regarding its quantity of output and selling price with a given product turns in part on the demand curve for the firm's output but it also depends on the amount which it costs to produce the product and on the relation of this production cost to the rate of output which the firm undertakes. We have surveyed the conditions of demand for the individual firm's, and the industry's, output, and we have noted how they are determined in a given state of total money purchasing power. The next step is to examine the determinants of costs and of the relation of cost to output.

The most important thing in this regard is, of course, the general level of cost—the interval within which cost per unit of output will fall over any probable range of outputs. What determines the costliness of producing a good? The cost in terms of money will evidently depend upon (1) the amount of productive services—labor hours, machine hours, and so forth—used in producing a unit of the product, and (2) the money prices of these productive services—the wage rates, prices of machines, etc. The amount of productive services used are the *real costs* of the good; the total money outlay required to purchase these services at their market prices is its *money cost.* The general level of the real cost of production so defined depends first on the *product;* it varies greatly from one good to another, depend-

ing upon the size and complexity of the good. Thus automobiles are quite costly in real terms; refrigerators require fewer man and machine hours and materials; hairpins are much less costly. But the real costs depend also on (1) the technique of production employed, as reflected in the types and design of the productive plant, and (2) the efficiency of the productive factors employed. The real costs of production will vary as technique or type of plant is varied, and also as the efficiency of labor and the quality of machinery change. The general level of the real cost of production is thus defined when we know (1) the product—as to type, design, quality, etc., (2) the technique of production and type and design of plant, and (3) the prevailing level of efficiency of employed productive factors. The last may frequently be given to the firm as outside its control, but the firm will continually make decisions regarding product and technique in order to get the best adjustment of cost to demand. Given its choices in these regards, it has a certain general level of real costs of production. If we know in addition the money prices of the productive services it employs we have also the general level of its money costs of production. It is this general level of money costs, and its relation to the general magnitude of demand at various money prices, which is really most important in determining whether or not a good will be produced and in what quantity it will be produced.

The level of cost, however, is not given and fixed regardless of the rate of output at which the firm produces. It will vary also with (1) the size or scale of the productive plant and of the firm operating a plant or plants, and (2) with the rate of utilization, or percentage of full capacity used, of the plant. The firm may build to different scales or capacities; having attained any given scale, it may utilize various proportions of its capacity. Changes in either scale or rate of utilization will potentially influence real and money cost. In effect, *variations in scale* and *variations in rate of use* of plant or firm are two methods of getting *variations in output;* output variations accomplished in either way will influence cost of production. It is apparent, then, that the firm will be concerned with the relation of cost variation to output variation, and in particular to output variation as

effected in each of the two ways. For decisions affecting relatively "*short periods*" over which the size of plant cannot be significantly changed, it will be concerned with the relation of cost to variation of output within the existing plant, or to variations in rate of use of this plant. For decisions affecting "*long periods*" over which the size of the plant can also be varied, it will be concerned with the relation of cost to variation of output accomplished by simultaneous variations in the scale *and* the rate of use of the plant. These relationships of cost to output are fundamental in the firm's determination of the most profitable rate of output.

This is not to say that the firm makes decisions only with respect to the rate of output with given product and given technique of production. It must at least intermittently decide exactly what product to produce and what technique or design of plant to use, and this with an eye to the effect of product and of technique on the cost of production. Given these decisions, however, and given also the prevailing efficiency of productive factors and their money prices (both ordinarily outside the firm's direct control), it will be positively concerned with the relationship of the money cost of production to the rate of output.

THE DEFINITION OF "COST"

Before analyzing this relationship, we should insert a word on the definition of *cost*. The money cost of a certain output as viewed by the average firm will be measured by the monetary expense paid or contracted for in obtaining that output; and this in turn will be that proportion of all outlays made by the enterprise (past, present, or future) which the firm considers allocable to the production of that output. This money cost will in turn represent the allocable real costs, in terms of productive services, times the money prices the firm has paid for them. Properly, the firm should also include in its costs not only *payments* made for goods and services but also the ~~*imputed value*~~ (or "opportunity price") of the productive factors which it uses but does not directly pay for. These costs would include the imputed but unpaid wages of an owner-manager, or the im-

puted interest charge on stockholders' investment, ordinarily not shown in accounts. For analytical purposes, then, the contractual or accounting costs of the firm should be considered as appropriately rectified to register the full opportunity prices of all productive factors employed. In general, this means that all labor, capital goods, materials, land and resources, and managerial service used by the firm in production should enter into costs at their respective market values regardless of the specific amounts paid by the firm in question. It also implies that full costs include a "normal" return, or free market reward, to all factors, including owners' investment and management, *and thus include a "normal profit" to the firm.*

Costs being thus defined, the relationship of cost to output (with given product, technique, money prices of productive factors, and efficiency) can be represented in a *cost schedule* for the firm, which shows the alternative costs of production at which various alternative outputs can be produced. The same information can be shown diagrammatically in a *cost curve,* which plots the variation of cost with output. Such a cost schedule or cost curve can be either for the *short run* (assuming a given fixed plant), or for the *long run* (assuming the size of the plant to be variable). In either case, the cost curve represents the *net relation* of cost c to output q. Generally cost depends upon or varies not only with output q, but also with money factor prices, p_1, p_2, etc., and with other variables. The conventional cost curve shows the net relation of c to q when all other related variables are held constant at certain levels. It does so legitimately because the money factor prices and other variables will not ordinarily vary *in response* to variation in the firm's output, and thus do not thereby influence the net relation of cost to output. (They may of course vary independently, and such variation, as we will see, may cause a shift in the cost curve relating cost and quantity.) If there is a systematic response of factor prices to the firm's output, however, the cost curve should reflect the resulting effect on aggregate costs; the special variety of cost curve which reflects such a response will be discussed in Chapter 7. We will first investigate the conventional *short-run* relation of cost to output for a typical firm, and then turn to the corresponding *long-run* relation.

DEFINITION OF THE "SHORT PERIOD" AND OF FIXED AND VARIABLE COSTS

This "short-period" relation of cost to output necessarily refers to a somewhat arbitrarily defined time interval. Ordinarily we consider it an interval during which certain productive factors employed by the firm, such as the building, heavy machinery, and permanent supervisory staff, are present in fixed or invariant amounts, and during which the quantity of other factors, such as labor and materials, is potentially variable in amount. Yet there is no clear-cut distinction in practice between "fixed" and "variable" factors. In effect, the longer a time period we contemplate, the more factors are potentially variable and the fewer are fixed. In formal logic, therefore, the short period to which we will refer is no especial chronological time interval at all, but a sort of "operational period," arbitrarily defined as of such length that buildings and long-lived equipment are invariant in quantity, and that labor and materials are freely variable in quantity. The common-sense counterpart of this interval for most manufacturing enterprises would perhaps be from six months to two or three years from any beginning date. For other types of enterprise, it might be longer or shorter.

In this short period, the firm has (by definition) certain "fixed factors" and certain "variable factors"—plant and equipment, let us say, on the one hand, and labor and materials on the other. It can increase or decrease its output in this period by varying the amount of variable factors it uses, or, in effect, by varying the *proportion* between fixed and variable factors. Correspondingly, the firm in the short period finds that its costs fall into two general categories—*fixed costs* and *variable costs*. *Fixed costs* are those which in the short peiod are absolutely invariant to changes in output; in precise terms, they are the amount of costs the firm would incur at a *zero* output. The *variable costs* are costs which vary with output, or, in effect, any costs added as a result of any increase of output above zero. In general the fixed costs will also be the costs of fixed factors, and will include depreciation of plant, interest cost on investment, and salaries of permanent managerial staff. Similarly, the

variable costs will ordinarily be the costs of variable factors, such as wages and material costs. But the categories are not necessarily fully congruent, and there may be some fixed-factor costs which are variable and some variable-factor costs which are fixed. The distinction between fixed costs and variable costs is therefore an independent one.

The content of fixed costs for a chosen interval is sometimes obscure, and an additional comment may be in order. Fixed costs for an interval will ordinarily include, first, costs incurred in the past—before this period—and allocated to the period. These are amortizations of past costs, like depreciation of the cost of equipment on hand which will occur even at zero output, although such allocations may be intrinsically arbitrary. Such costs are already "sunk" and cannot be lessened by any strategem. Second, fixed costs will include current outlays, made during this period, which will in any event be made at zero output. It will be noted that the second category of fixed cost, though fixed for the current short period, are evidently made in anticipation of operations in a future period. Although currently fixed, therefore, they are—for a *longer* period of calculation—essentially variable costs the future recovery of which is anticipated. It may also be noted that economic fixed costs and accounting overhead costs are not identical. The latter include any costs which are allocated by formula against different lots or units of output, and may include variable as well as fixed costs.

SHORT-RUN RELATION OF COST TO OUTPUT

Given these general considerations, what regularities can be observed in the short-period relation of cost to output for a business firm? The cost schedule shown on the next page indicates a pattern which is usually approximated. The first column shows various alternative outputs, and the second the aggregate cost of producing each. If such a cost schedule is translated diagrammatically into a continuous curve (called an *aggregate cost curve*) it looks something like that in Figure 8.

This schedule and curve possess several characteristics which are felt to be typical of short-term cost variation for most firms.

Output	Cost
0	$ 50
5	58
10	65
15	71
20	76
25	81
30	87
35	94
40	102
45	111
50	121
55	132
60	144
65	157
70	171
75	186
80	202

exists the cost curve

First, as we have observed already, there is some positive fixed cost at the zero output level—thus the aggregate cost curve has a positive origin on the vertical axis. Second, the aggregate cost increases with increasing output. This is obvious and is evidenced by the fact that the aggregate cost curve slopes continually upward to the right. Third, the aggregate cost increases *at varying rates* in response to given variations in output; specifically, it first increases *at a decreasing rate*, and then begins to increase *at an increasing rate* as output is progressively extended. This is evidenced by the fact that the aggregate cost curve becomes progressively less steep in slope up to the output of 20, but becomes progressively steeper in slope after the output of 25 is passed.

This third property of the short-run cost curve is the least obvious and at the same time the most important. If aggregate cost increases first at a decreasing and then at an increasing rate, it is evident (1) that there is a corresponding initial increase and succeeding decrease in efficiency; (2) that there is some intermediate output where the *costs per unit of output* will be a minimum; and (3) that short of this output costs per unit will decline, whereas past this output costs per unit will rise with increasing output. This last generalization, as applied to any

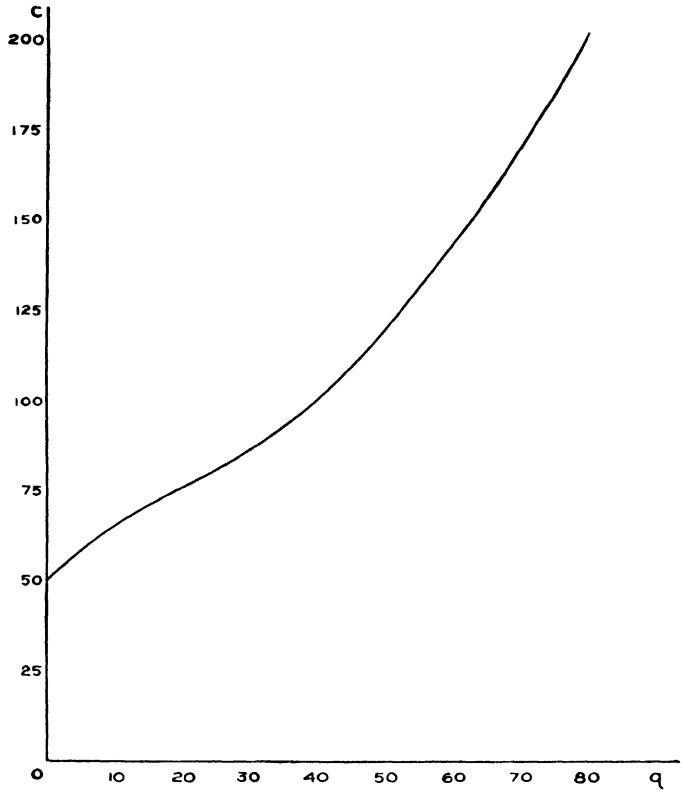

Figure 8

operation where certain variable factors are employed in conjunction with certain fixed factors, constitutes a "law" of varying costs. In effect, as the output produced from a given aggregation of fixed factors is successively increased, the costs per unit of output will at first decline and will then progressively increase.

The reason for this law of costs is not found in any variation in the price the firm pays per unit of variable factors, since this we assume to be constant. It stems rather from the variation in the proportions of variable and fixed factors which occurs as output is extended within a fixed plant, and which alters the physical productivity of the factors employed. As output in-

creases from zero, the proportion of variable to fixed factors becomes progressively greater. At some point we strike the proportion where costs are a minimum. Short of this point the proportion is becoming more economical and unit costs are falling; past this point the proportion becomes less economical and unit costs increase. The law of varying costs is thus based on a physical law that variations in the proportions among productive factors produce variations in physical efficiency.[1]

The preceding generalization concerning cost variations applies specifically where some productive factor is either invariant or less than perfectly variable in amount. If all factors were freely variable, an optimum and unchanging proportion among them could be maintained at any rate of output, and cost would increase in exact proportion with output.[2]

The significance of the typical pattern of short-run cost variation just described is more easily seen if we analyze the aggregate cost variation as follows:

1. By dividing the aggregate total cost schedule, already shown, into two main component schedules: *aggregate fixed cost* and *aggregate variable cost.*
2. By expressing aggregate fixed, aggregate variable, and aggregate total costs as *averages* per unit of output.
3. By calculating the *increment* in cost for each increment in output.

[1] The basic physical law was traditionally referred to as the "law of diminishing returns," but has been more frequently referred to by the less revealing title of the "law of variable proportions." In either event, the basic phenomenon referred to is this: If in producing any good certain factors of production are employed in fixed quantity and certain other factors of production are employed in varying quantity, successive equal increments in the quantity of the variable factors will be rewarded at first by increasing increments in output and then by decreasing increments in output. This tendency is the basis of successively decreasing and increasing costs per unit of output. See Kenneth E. Boulding, *Economic Analysis* (New York, Harper & Brothers, 1941), Chap. 22, for a detailed discussion of the laws of return and the derivation of cost curves.

[2] For this to hold, however, all factors would have to be perfectly divisible as well as freely variable (see pp. 86-88 below). And the "firm," or management coordination, should be variable without disadvantage. It is unlikely that both of these conditions would ordinarily be found in fact.

The results of those calculations, applied to the cost data from page 67, are expressed in the table on page 71.

Careful examination of this table reveals the principal properties inherent in the sort of aggregate cost variation we have characterized as typical. The following points may be noted:

a. *Aggregate fixed cost,* as shown in column (3), is, of course, constant for all levels of output—in this case at $50.

b. *Aggregate variable cost,* as shown in column (4), is calculated as the difference between aggregate total cost and aggregate fixed cost (column 2 minus column 3). This aggregate variable cost contains all the variation to which aggregate costs are subject.

c. *Average fixed cost,* or fixed cost per unit of output (column 6), is calculated by dividing each aggregate fixed cost (from column 3) by the corresponding output (column 1). Average fixed cost declines monotonically, of course, as a constant amount is spread over more and more output.

d. *Average variable cost,* or variable cost per unit of output (column 7), is calculated by dividing each aggregate variable cost (column 4) by the corresponding output (column 1). Average variable cost at first declines with increasing output, reaching a low of $1.23 per unit at the output of 30, and thereafter increases with increasing output. This variation reflects the general law of varying costs already referred to.

e. *Average total cost,* or total cost per unit of output (column 5), is calculated either by adding the average variable and average fixed cost at each level of output (column 6 plus column 7), or by dividing each aggregate total cost (column 2) by the corresponding output (column 1). It will be noted that average total cost also first declines with increasing output, reaches a minimum, and then increases. The minimum average total cost, $2.40, is reached at a larger output than the minimum average variable cost—at an output of 55 instead of at 30—because of the influence of declining average fixed costs. The variation in average total cost thus reflects the same law as variable

SHORT-RUN COST VARIATION

(1) Output	(2) Aggregate total cost	(3) Aggregate fixed cost	(4) Aggregate variable cost	(5) Average total cost *	(6) Average fixed cost *	(7) Average variable cost *	(8) Increment in cost (per unit of output)
0	$ 50	$50	$ 0				
							1.60
5	58	50	8	$11.60	$10.00	$1.60	
							1.40
10	65	50	15	6.50	5.00	1.50	
							1.20
15	71	50	21	4.73	3.33	1.40	
							1.00
20	76	50	26	3.80	2.50	1.30	
							1.00
25	81	50	31	3.24	2.00	1.24	
							1.20
30	87	50	37	2.90	1.67	1.23	
							1.40
35	94	50	44	2.69	1.43	1.26	
							1.60
40	102	50	52	2.55	1.25	1.30	
							1.80
45	111	50	61	2.47	1.11	1.36	
							2.00
50	121	50	71	2.42	1.00	1.42	
							2.20
55	132	50	82	2.40	.91	1.49	
							2.40
60	144	50	94	2.40	.83	1.57	
							2.60
65	157	50	107	2.42	.77	1.65	
							2.80
70	171	50	121	2.44	.71	1.73	
							3.00
75	186	50	136	2.48	.67	1.81	
							3.20
80	202	50	152	2.53	.63	1.90	

*Averages are computed only to the nearest cent. This obscures some details of the average cost variations.

cost but is complicated by the variation in average fixed cost.

f. The increment in cost per unit of output is shown in column 8. This represents the increase in cost for each increase in output, expressed on a per-unit basis. Thus, when the output moves from 0 to 5, the increment in cost is $8 for 5 units, or $1.60 per unit; when output moves from 5 to 10 units, the increment in cost is $7, or $1.40 per unit. Following the law of successive decrease and increase of cost, the incremental costs also at first decrease and then increase.

It will be noted that the incremental cost is necessarily less than average variable cost while the latter is falling, and greater than average variable cost while it is rising. It bears the same relationship to average total cost. If output is varied continuously rather than by discrete increments, incremental cost is necessarily equal to either of these averages at that output point where the average reaches its minimum. This is because each increment in cost enters into and influences (either upward or downward) the preceding average cost.

Viewed as a whole, this tabular breakdown of a typical aggregate cost variation makes clear the structure of costs for a firm in the short period. The firm finds that its costs vary in response to variations in output in such fashion that, as output is extended, variable costs per unit of output at first decline and then increase, ordinarily at an increasing rate. Similarly, the increment to cost per unit of output first declines and then progressively increases as output is increased. Average fixed costs decline with increasing output, but at a decreasing rate, as the fixed "overhead" is spread over more and more units of output. After the output for which average variable costs are a minimum is exceeded, average total costs will still fall for an interval with increasing output, because the decline of average fixed costs will exceed the rise in average variable costs. But as some rate of output is reached, the rise of average variable costs just counterbalances the dwindling decline of average fixed costs, and here the minimum average total cost is reached. This output for which average total cost is a minimum has been called

short-run optimum output for the firm, or optimum rate of utilization for its given fixed plant. (It is, indeed, the output which allows of lowest unit costs for a given short period. As we will see below, however, this output may not be especially sought after by the firm, and only under certain fortuitous circumstances will it be the most desirable output from a social standpoint.)

The cost relationships just described may be illustrated more clearly by using a diagram, such as Figure 9, in which each cost component is expressed as a continuous curve. Thus for the numbers in the average-fixed-cost schedule we substitute a continuous curve, *AFC*, which shows the level of average fixed cost for every possible graduation of output. Similarly, average variable cost is shown as a continuous curve, *AVC*, as is average total cost, *ATC*. In place of the schedule of cost increments, we substitute a *marginal cost curve, which shows the rate of increment of total cost per increment in output.* All these curves are derived directly from the original aggregate cost data shown in the table.[3]

The cost variations shown in the table appear about as in Figure 9 when translated into continuous cost curves. The pattern of cost variation thus illustrated has, as we have indicated before, several characteristics which are thought to be typical of short-run cost curves generally. Diagrammatically, these appear as (1) an average-variable-cost curve which has some sort of a U or "dish" shape, falling to a minimum and then rising; (2) an average-fixed-cost curve which declines monotonically at a progressively decreasing rate; and (3) an average-total-cost curve which, as the sum of the first two curves, also is U or "dish" shaped, declining to a minimum (at an output greater than that for which average variable cost is minimized) and then rising. The marginal-cost curve typically falls for initial increases in output and thereafter rises continually.

[3] Thus let TC = aggregate total cost; VC = aggregate variable cost; FC = aggregate fixed cost; X = output. Of course $TC = FC + VC$. Also $ATC = TC/X$; $AVC = VC/X$; $AFC = FC/X$; and $AVC + AFC = ATC$. Further $MC = \frac{d(TC)}{dX}$ = the rate of change of TC in response to change in X.

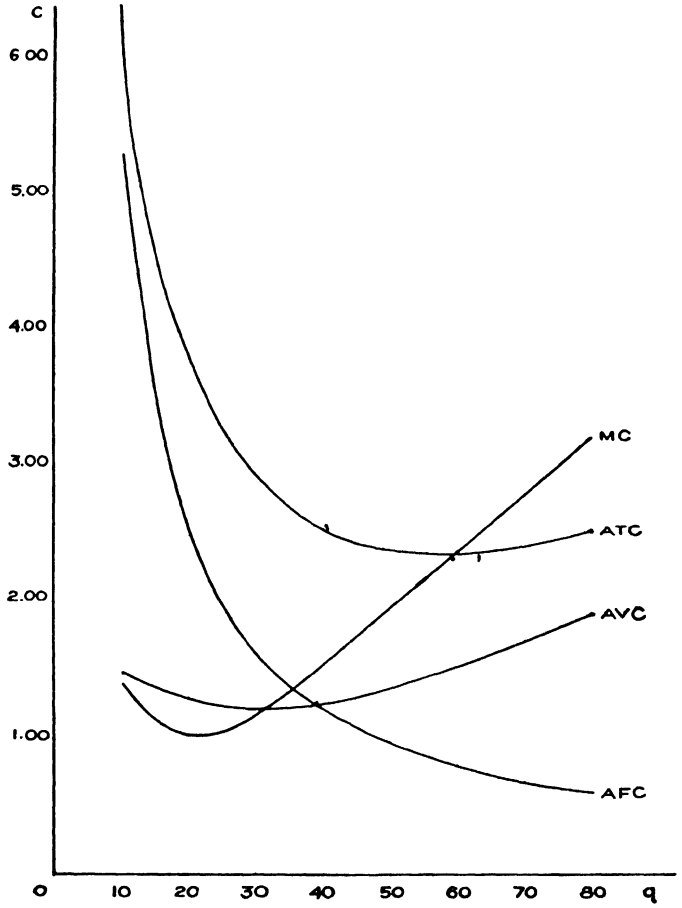

Figure 9

The mechanical formal properties of such a family of cost curves are clear from our discussion above. The curve *MC* necessarily intersects *AVC* and *ATC* at their respective minima; the minimum of *ATC* lies at a larger output than the minimum of *AVC*; the curve *AFC* is in every case of a given fixed shape—a so-called rectangular hyperbola. It is also clear that the rising marginal cost and U-shaped average-variable and average-total-cost curves reflect the varying proportions of variable to fixed factors as output is progressively increased.

The full significance of these cost variations to an enterprise must be explored in succeeding chapters. It should be clear, however, that the cost curves are drawn to summarize certain information of importance to the firm: (1) the movement of total cost per unit of output with variation in output (*ATC*); (2) the corresponding variation of average variable costs (*AVC*); (3) the corresponding variation of average "overhead" or fixed cost (*AFC*); and (4) the *additions* to cost for given additions to output (*MC*). Considering such variations in conjunction with the price variations shown by his demand curve, the entrepreneur should be able to select the price or output which will maximize his profit.

This representation of the variation of cost with output in the short run and with given plant is drawn on the supposition that the prices of the productive factors which the firm employs are given at certain levels and do not vary by reason of variations in the firm's output. Wage rates, for example, are $1.50 per hour regardless of output, and the same is true of material prices and machine costs. This is because the firm's employment of factors is supposed to be so small that variations in it will not influence factor prices. As a consequence, the variation in average and marginal costs shown reflects only varying efficiency due to varying proportions of fixed to variable factors, and *not* to any variations in wage rates or other factor prices. Where the prices of factors vary systematically because of variations in the firm's output, the behavior of the cost curve is more complicated. Such situations will be discussed in Chapter 7.

The preceding cost-curve diagram, however, does not purport to describe exactly in all respects the short-run cost variation experienced by all firms and industries. It is typical in that it reflects (1) the inclusion of at least some fixed cost, and (2) the incidence of at least some upward variation in marginal and variable costs as output increases beyond a certain point. From firm to firm, and more particularly from industry to industry, there are very significant differences in (1) the proportion of variable to fixed costs, and (2) the shape of the variation in average variable and marginal costs. In the example we have given, for instance, the proportion of fixed to variable costs is fairly high, and the variation of average variable and of mar-

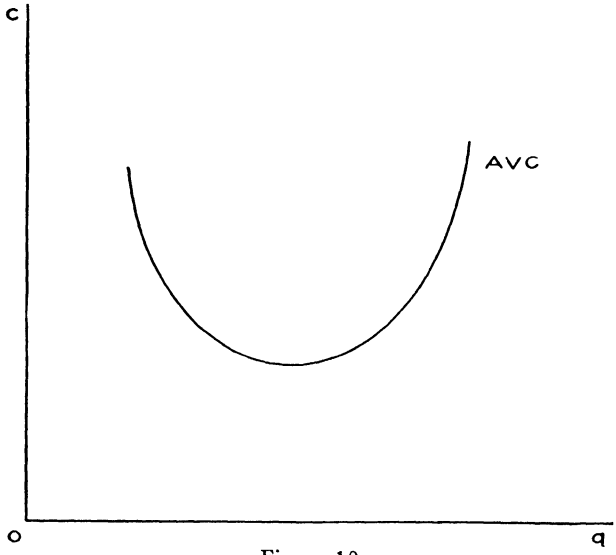

Figure 10

ginal cost is comparatively gradual (although at some output larger than we have shown average variable and marginal costs would necessarily rise steeply as the physical capacity of the plant was reached). This gradual variation of costs, reflected in a rather flat or shallow U shape in the average-variable and average-total-cost curves, and in an only moderately sloping marginal cost, indicates that the plant in question could be used fairly efficiently over a rather wide range of output—the *rate of utilization* of plant could be varied widely with rather moderate variations in unit cost.

Cost variation for firms in other industries, however, might be very different. Textbook examples often favor a family of cost curves which shows the average variable cost curve with a decided U shape, as in Figure 10. Here, average variable costs rise rather steeply as output moves in either direction from the optimum, so that there is only a small range in the rate of plant utilization which gives reasonable efficiency. Such cost behavior might characterize plants with certain technical peculiarities. But it is equally possible that the pattern of cost variation may, within the general limits of the law of varying costs, fall any-

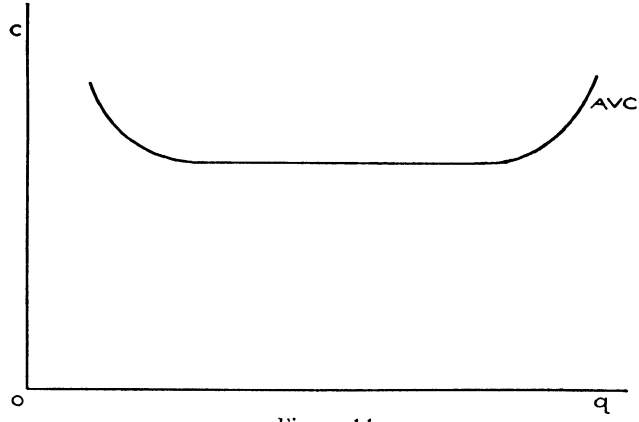

Figure 11

where between wide extremes. At one extreme the average variable cost curve may be a very wide and flat-bottomed U, showing no cost variation over a considerable range of output, as in Figure 11. Here, the rate of utilization could be varied from possibly 20 percent to 90 percent of maximum capacity without influencing variable costs very much.

At another extreme, the average variable cost curve may be practically V-shaped, as in Figure 12. This would indicate that only at one critical rate of utilization was any reasonable approach made to the attainable efficiency of the plant. Various instances found, in fact, may lie anywhere between those extremes, and may also show peculiar variations of their own.

The reason for substantial differences in cost behavior among different firms is evident in the basic reason for short-run cost variation. In the short-run variable and marginal costs at first fall and then rise with increasing output because of the changing proportions of variable to fixed factors—of labor and raw materials to plant. Inherent in the technical design of any plant will be a ratio or coefficient between labor and materials on one hand and machinery and equipment on the other which allows optimum unit costs. Production at such a rate that this ideal ratio is just maintained should give minimum costs; production at lower or higher rates will cause costs to increase. But the degree to which costs respond to variations in output may differ

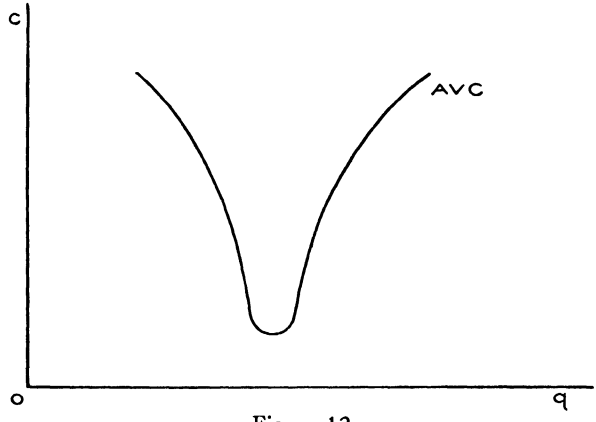

Figure 12

widely among plants. In the first place, the technical character of plants may differ considerably. In some plants, the ratio of variable to fixed factors may be technically inflexible, so that output cannot be had at all without a given "dose" of labor and materials per plant unit and cannot be increased much by adding more variable factors per plant unit. This might be true, for example, of a mechanical punch press operated by one man—it needs one man and only one, and consumes no less or more than so much material per hour. In other plants, the ratio of variable to fixed factors may be technically variable away from their optimum proportion with only a moderate disadvantage in cost. Thus a road-building company with given fixed equipment might employ from one to several men (up to the final limit of overcrowding) with only moderate variations in variable unit cost. The more variability in the ratio of variable to fixed factors that the technique and design of a plant allow, the shallower and flatter the U-shaped variable-cost curve will tend to be.

A more potent consideration affecting the pattern of short-run cost variations in a plant is the *divisibility* of the plant. If all of a plant must be operated at once in order to run at all, then employing a small proportion of the normal work force will greatly affect the working proportion between labor or other variable factors and fixed plant. But if the plant is divisible into identical fractions or parts, each of which may be operated

separately while the other parts are closed down, it will be possible to maintain an efficient proportion of variable to fixed factors while employing widely different amounts of variable factors. Thus if a steel plant consists of a single technical unit—let us say one Bessemer converter—any considerable variations in work force or amount of materials used will seriously affect costs. But if an iron-making plant is made up of ten largely independent blast furnaces, any number of which can be operated at a time, the work force might vary from 10 percent to 100 percent of the maximum without greatly influencing the average variable cost of producing iron. In the latter case, the average variable cost curve might be almost horizontal over most of its length, with the U-shapedness appearing only at extremely small or large outputs, as in Figure 11. In short, divisibility of the plant into numerous identical subdivisions tends to lessen cost variability, and to result in wide flat-bottomed U-shaped cost curves—up to capacity. Technical indivisibility of plant is associated with more distinct variability of short-run average variable costs.

Nonvariability of average costs is also favored where a plant may be operated for any fraction of the working week or month. A short-run cost curve shows variation in unit cost in response to variation in output per some unit of time, such as a month. In most plants, such a variation in the rate of output does not involve using more or fewer variable factors in a plant at any one time, but may be accomplished by operating more or fewer days per month, or more or fewer shifts per day. As long as the number of shifts worked per week or month is easily variable, the rate of output per month can be varied widely without affecting very much the physical proportion of labor to plant, and therefore without much influencing average variable costs. Wherever the technique and other governing conditions are such as to allow this variation in time worked, therefore, we tend to find relatively constant average variable costs over a wide range of output, reflected in a flat-bottomed U-shaped cost curve. Where the technique does not easily allow alternate shutdowns and start-ups of plant, costs will tend to vary more with variations in output.

In short, although any enterprise operating with fixed plant will experience some variation of average costs in response to varying output, and although unit costs will ultimately rise as physical capacity is reached, the degree of variation in average variable and marginal costs may differ greatly from case to case. In some cases, where intermittent plant operation is not feasible and where the plant is an indivisible technical unit, the average-variable-cost curve may be a very steep U, with unit costs falling greatly as optimum utilization is approached and rising quickly as it is surpassed. In many others, where the plant characteristics are otherwise, average variable cost will tend to be relatively constant over a wide range of output, and significant cost variations will occur only at extreme outputs.

In American manufacturing industry, where plants are often very large, where a single enterprise may operate several similar plants, and where intermittent operation is often feasible, the phenomenon of *constant average variable costs* (except at extreme outputs) has been frequently noted. For this type of business enterprise, at least, statistical studies lead us to believe that the typical pattern of cost variation is represented in a wide flat-bottomed U shape for the average-variable-cost curve, and in constant or unvarying marginal costs over a wide range of output. This may be significant in price determination.[4]

Although the variability of average variable costs in the short run is of primary importance, considerable significance attaches to the relative size of fixed and variable costs. In certain industries fixed costs are very large, amounting to as much as 50 percent of the total unit cost at capacity output; in others they are relatively insignificant. The relative size of the aggregate fixed cost is important because fixed costs per unit of output always decline in a set pattern with the extension of output. If fixed cost is a very large proportion of total cost, this decline of average fixed cost is likely to dominate the whole pattern of cost variation; whether variable costs are constant or sharply rising,

[4] For an example of the empirical study of cost behavior, see Joel Dean, *The Relation of Cost to Output for a Leather Belt Shop* (New York, National Bureau of Economic Research, 1941). A general discussion of cost behavior appears in National Bureau of Economic Research, *Cost Behavior and Price Policy* (New York, 1943).

the average *total* cost of production will fall significantly over a relatively wide range of output. In a plant with big fixed costs, great economies attach to full production because of the connected opportunities for "spreading overhead" over large outputs. In a plant with small fixed costs, the pattern of variation of variable costs will dominate the variation of average total cost.

Our discussion of costs so far has centered about the variation of production costs in response to variation in output, where the output variation is effected by varying the rate of utilization of a given fixed plant. We have thus emphasized the short-run cost variation with which a firm is concerned in making decisions over time periods which are too short to permit it to vary the size of its plant. This so-called short-run cost variation is thus ordinarily relevant to price-output decisions which affect the next month, six months, or year. Before another type of cost variation is discussed, two additional comments on the short-run cost curves may be added.

First, it should be re-emphasized that the characteristic pattern of short-run cost variations discussed above, which features a successive fall and rise in average variable costs and in uninterrupted decline in average fixed costs as output is increased, is a reflection of the existence of fixed factors which are invariant in amount over the period for which output variations are contemplated. The existence of such factors gives rise not only to fixed cost, but also to the varying proportionality of factors which accounts for changing average variable costs. The short-run cost pattern is not characteristic of situations where all factors are variable and where the proportions of all factors can thus be kept constant at different levels of output.

Second, the formal properties of short-run cost curves are such that they show the *net relationship* between variation in output and variation in costs. Movements along the cost curve show only those changes in costs which occur directly in response to changes in output; they do not show cost changes which are due to independent changes in wage rates or material prices, or due to independent variations in the quality of factors. Since the typical assumption is that the latter variables will not vary in response to the firm's output variations, but only independently, the cost curve is typically drawn on the assumption that wage

rates, etc., are constant, as above. We consider cost c, as one variable, which is specifically related to several other variables, including output q, the wage rate W, material prices M, and so forth. Cost may vary in response to changes in q or W or M or all of them. The cost curve, however, typically shows only the net relationship of c to q, assuming that all other variables which influence c remain constant as output varies. *Variations in cost which are caused by independent changes in wage rates and material prices are shown not in movements along our cost curves but in shifts of these curves.* In the less typical case where wage rates and material prices vary systematically in response to variations in the firm's output, the effect of such connected variations is reflected in movements along average and marginal curves. (See Chapter 7.) But any *independent* variation in factor prices is still reflected in shifts in the curves.

COST VARIATION IN THE LONG RUN

Short-run cost calculations, since they show how a firm's costs will vary in response to variation in output within the limits of a given fixed plant, are relevant so long as the firm is concerned with a short future time period within which it cannot greatly expand or contract its plant. But the firm will also be concerned with the relationship of cost to output for successively longer periods, including intervals long enough for it to vary the size of its plant freely. A simplified version of this relationship is the theoretical "long-run" relation of cost to output, calculated on the assumption that there are no fixed factors, or no given size of plant.

The *long-run cost curve* of a firm is correspondingly one which shows the variation of cost in response to variation in output for a period long enough that all factors of production, including plant and equipment, are freely variable in amount. The "long" period in question, like the "short" period, thus has no certain chronological limits but is functionally defined. But it is fairly approximated in intervals of three to ten years over which firms may calculate their plans for long-run expansion (or contraction).

In precisely what decision-making context would a firm have reference to such a long-run cost curve? It will refer to it in connection with a period which (1) begins far enough from now that the firm will have time fully to adjust the size of its plant to the necessities of that period, and (2) lasts long enough after its beginning that all durable goods acquired for use in the period can be fully used up or worn out during the period and that their costs can be fully amortized in accordance with plan. The long-run cost curve, appropriately referred to as a "planning curve," is applicable in deciding the relation of the firm's scale of plant to an average situation of demand for its output over a future period of considerable length.

A long-run cost curve for any firm, like a short-run curve, shows the net relation of cost to output. Like a short-run curve, it can represent average costs or marginal costs. The principal formal difference is that *in the long run there are no fixed costs.* All costs are variable. The family of cost curves will therefore consist only of an average-total-cost curve and a marginal-cost curve. If long-run average costs followed the same sort of U-shaped variations as short-run average costs, therefore, the long-run average and marginal costs of a firm would appear as in Figure 13. The absence of any distinction between fixed and variable costs will be noted.

In drawing any long-run average cost curve, like that illustrated, it is supposed that the plant, equipment, and all other factors are freely variable. At the same time, the rate of utilization of any chosen plant is also freely variable. The drawing of the long-run average cost curve thus assumes that the scale of plant and the rate of use of any plant are coordinated and simultaneously adjusted so as to produce any chosen output at the lowest possible cost. Any cost amount from a long-run average cost curve which corresponds to a specific output is thus the lowest cost at which that output can be produced, *scale and rate of utilization of plant being coordinately chosen.*

Does the long-run average cost curve also have a typical shape, which is found with some uniformity in all sorts of business enterprise? The answer to this question is not entirely certain, although investigations of the relation of cost to scale in a large number of industries give us some basis for generaliza-

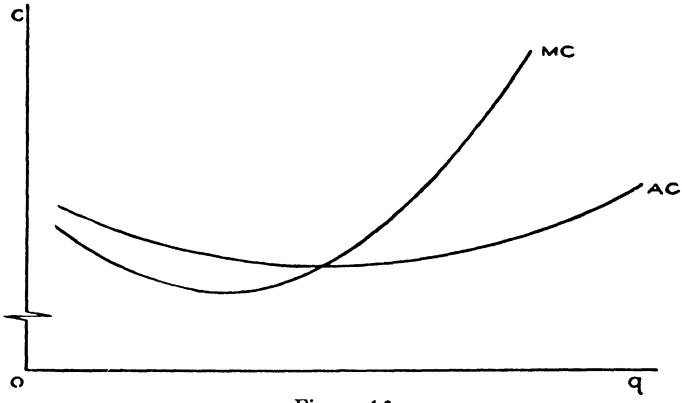

Figure 13

tion. Speaking very broadly, it is possible to say that the long-run average cost curve is typically U-shaped—that as *the scale of the firm* is progressively extended, the cost per unit of output at first declines, then becomes constant, and finally increases. Thus the "typical" long-run average cost curve has the U or "dish" shape illustrated in Figure 13.

Empirical studies of industrial costs seem to support this statement. In most industries a very small firm is quite inefficient; as the firm becomes larger it tends to become more efficient, reaching a minimum cost per unit of output at some particular scale. Thereafter, further expansion up to some still larger scale may have little effect on unit costs, which remain constant with increases in capacity. But if the scale of the firm is made still larger, a point ordinarily seems to be reached where unit costs begin to rise as the firm becomes "too large."

The explanation of this phenomenon in terms of the common sense of engineering and organization has been extensively developed. It is suggested that in most industries there are certain *economies*, or cost savings, available through expansion of scale up to a certain point. These include the economies of greater *specialization of labor*, of greater *specialization of capital equipment*, of *large* as opposed to *small mechanical devices*, of *large-scale distribution*, and so forth. As a firm becomes larger it can reduce its average costs by taking advantage of those economies.

(We refer strictly to "real" economies of using factors priced at given levels; separate attention may be given to pecuniary economies due to any decline of factor prices with increasing scale.) The range of alternative scales over which it can so reduce average costs is referred to as a stage of *decreasing costs* in response to increasing scale. The extent to which further and further expansion of scale will result in falling average costs will differ from industry to industry according to the sorts of technique they employ. After scale has been extended to some critical point in each case, however, the economies of increasing scale will have all been fully exploited, and further additions to the size of the firm will not result in further reduction of unit cost. Since plants of optimum size can be duplicated by a single firm, on the other hand, there is apparent no automatic tendency for costs to rise with still greater scale. Once all economies of increasing scale have been exploited, costs should tend to remain constant in the face of further expansion, unless some other factor intervenes.

The reason that average costs eventually tend to rise with further increases in scale, it is held, is found in just such an additional factor—the *diseconomies* of very large-scale management. If a firm becomes large enough, the argument goes, the burden of administration becomes disproportionately great; "red tape" tends to proliferate; and total costs per unit of output tend to rise. Thus we get a terminal stage of increasing costs with increasing scale which adds the last leg to the U-shaped long-run average cost curve.

All the preceding is probably true in a general sense as applied to many firms and industries. It should be emphasized, however, that within this general pattern there are big differences in long-run cost behavior among industries. In certain industries, economies of increasing scale are realized for expansion of the firm up to such a point that the whole market would be fully supplied by a single firm before it had reached its optimum size. Such an industry may be called a "natural monopoly," and the firm in question may never grow big enough to encounter the ultimate upturn in its long-run average cost curve. In other industries the optimum scale of firms may be large enough that only a few firms of optimum size

can be supported by the market. In still others, a very small firm may be able to exploit the economies of scale fully, and the market may be able to support many firms of the most efficient size. One of the reasons that in so many of our industries there are very few firms is that in most manufacture the economies of large-scale mass production are great and efficiency is increased by concentrating production in a few hands.

The variation of a firm's production costs in response to varying output in the long run, in sum, tends to be such that the cost curve relating average cost to output is of U shape, evidencing successive fall and rise of unit cost as output is extended. Correspondingly, the firm's long-run marginal cost curve, showing increments in cost with increasing output, will usually first fall and then rise, always intersecting the average-cost curve at its lowest point, as in Figure 13.

In terms of such a relationship of unit cost to rate of output (considered in conjunction with the demand for its product), any business firm will decide how large to make its long-run scale of operations. We may anticipate here a later discussion by remarking first that it is not at all inevitable that the firm will select the "optimum scale"—that is, the long-run rate of output which gives the lowest unit cost—and second that it is only under special conditions of market structure that production by firms at exactly this "optimum" scale will be most desirable from the standpoint of society. From a social standpoint it is important that the output of any industry be related to that of other industries in an "ideal" fashion (yet to be defined), and that the resulting industry output be produced at the lowest attainable aggregate cost. If this output cannot conceivably be produced with firms operating at optimum scales, then any necessary departure from these scales is consistent with maximum social welfare.

The rule that the usual firm's cost per unit of output is a minimum at some determinate scale of operations and becomes higher if the scale is made smaller or larger than this, has so far been supported only by referring to certain "economies of large-scale production" and to "diseconomies of very large-scale management." Although such economies and diseconomies are unquestionably found in practice, reference to them does not put

the explanation of varying long-run costs on the same footing as that of varying short-run costs. In fact, it leaves unsettled a clear implication of our earlier discussion to the effect that, with all factors freely variable, we might expect no variation at all in long-run average costs in response to varying scale.

In the discussion of the short period, it was pointed out that the tendency of short-run variable costs first to fall and then to rise as the rate of utilization of a fixed plant is increased results from the changing *proportion* of variable to fixed factors. The optimum rate of utilization can be struck at only one rate of output, where this proportion is the best. This is because, in the short run, part of the factors of production are fixed instead of being freely variable. But if all factors are freely variable in amount, as they are in the long run, it would seem that at any rate of output the ideal proportion of factors could be had, and that therefore the long-run rate of output or scale of operations should have no effect on unit costs. Unit costs should remain constant at a minimum level for any scale of operations. How may this logic be reconciled with our observations concerning the economies and diseconomies of increasing scale?

The key to this apparent dilemma is simply that there are concealed limitations on the free variability of all factors in the long run. Although all factors of production, excepting perhaps management, are freely variable in amount, *the various factors* (capital equipment, labor, and resources) *are not indefinitely divisible into small units.* Thus although it is possible for a firm to use more or less capital equipment within wide limits, it is not always possible to use more or less equipment in the specific forms it is needed. It must use a whole belt-conveyor for its assembly line, or none at all—it cannot use 1/100 of a belt conveyor. Similarly, although it can use more or less labor within wide limits, it cannot use the various specialized forms of labor in fractional units. A big plant may be able to use one production-control statistician to advantage; a small plant cannot ordinarily use 1/10 of a statistician. In brief, *the recognized economies of large-scale production are essentially a reflection of a technical indivisibility of factors of production within given patterns of technique.* This indivisibility prevents the attainment of the best proportion among factors until a scale of operations

large enough to overcome all indivisibilities is reached. The decline of unit costs in response to expansion of scale thus reflects a progressive approach to the best proportion of factors. Varying proportionality of factors is at the heart of long-run cost variation just as it is for short-run cost variation.[5]

The same line of argument holds in explaining the rise of unit costs as scale becomes too large. If diseconomies of large-scale management are encountered, this is because the management factor cannot be indefinitely expanded to match expansion in other factors; management is a "fixed," or more properly an "imperfectly variable," factor, and indefinite increase in the amount of other factors will thus result in rising costs. In other words, the "firm" itself, or its administration, is a concealed or passive fixed element which accounts for varying costs even when the proportions of all other factors may be varied at will. The U shape of the long-run cost curve is thus a reflection of varying factor proportionality with variations in the scale of the firm, although this variation has a somewhat different origin from the varying proportionality found in the short run.

Throughout the preceding discussion of long-run cost curves, we have assumed that the factor prices paid by the firm do not vary in response to variations in the firm's output, and thus constructed the curve on the assumption of given factor prices. The economies and diseconomies referred to are thus entirely technical or real economies and diseconomies, and they alone cause long-run costs to vary. Where the firm becomes very large, however, there may in addition be systematic variations in factor prices in response to variations in the firm's output, giving rise to strictly pecuniary economies or diseconomies of large scale. Where these are encountered, they should be reflected in the shape of the long-run cost curve. This special phenomenon is neglected below, but will be considered in Chapter 7.

To this point we have characterized as having in general a U shape the average-total-cost curves showing, respectively, short-run and long-run cost variations in response to varying output. One of these curves shows the firm how unit cost will respond to varying output within its present fixed plant; the

[5] Cf. Boulding, *op. cit.*, Chap. 23.

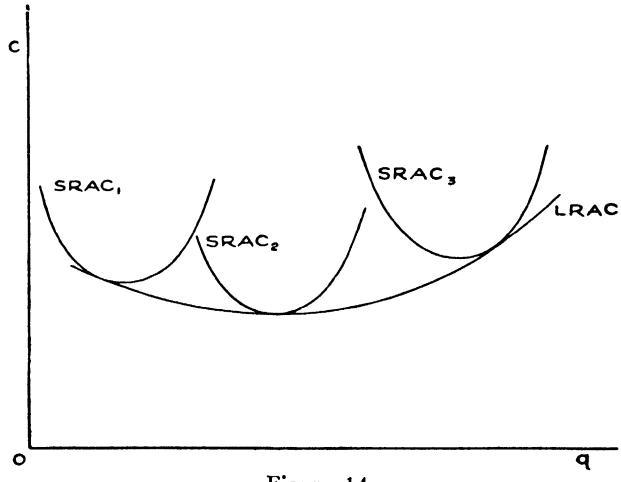

Figure 14

other shows it how unit cost will respond to varying output over a longer interval during which both size of plant and rate of use of plant can be varied. How are these two curves related to one another? Their formal relationship is defined in our earlier remark that the long-run average cost curve shows for each possible output the lowest cost of producing that output, size of plant and rate of use of plant being covaried to obtain the best result. If this is the case, *the long-run average cost curve is hypothetically an "envelope curve," which just skirts the lower edges of an indefinite succession of possible short-run cost curves.*

Thus, in Figure 14, $SRAC_1$, $SRAC_2$, and $SRAC_3$ represent the short-run average total cost curves attached to each of three successive scales of operation. The curve *LRAC* is the long-run average cost curve which generalizes the long-run relation of cost to output shown by the movement from one short-run curve to another. The U shape of the long-run curve implies that the short-run curves for successively larger plants have lower and lower average cost minima up to a certain "optimum" scale, and thereafter higher and higher average cost minima. Since we draw the long-run cost curve as a continuous line, we assume that plant may be increased in size by indefinitely small

gradations; every point on *LRAC* is thus a point of tangency with some short-run cost curve which it envelops.[6]

We have so far discussed the relation of cost of production to output for business firms and have discovered certain typical patterns of cost variation which are likely to influence all price determination. But in so doing we have centered attention on only one of the changes which can cause a firm's production costs to vary—namely, the variation in its output of a given good. As we have specifically indicated already, independent changes in wages or in the prices the firm pays for its materials are not reflected in our cost curves, although these changes can be readily represented by shifts in these curves. Thus the average and marginal cost curves of a firm when wages were $5 an hour and materials and equipment were at some specific price might appear as AC_1 and MC_1 in Figure 15. These curves typically show the variation of cost in response to change in output, given the stated level of wages and material prices and positing no response of wages and material prices to output. If wages rise, for example to $7 an hour, and materials to some corresponding higher level, we would register the resulting change in production costs by a shift of both cost curves, to the new positions AC_2 and MC_2. The new cost curves would show the relation of cost to output, given the new wage-price level. The effect of independently changing wages and material prices on costs is thus shown in the shift from one curve to another (and also in any change in the shapes of the cost curves, resulting from disproportionate changes in wages and material prices, which might

[6] It will be noted that the *LRAC* curve does not intersect the minimum-cost points of any of the short-run curves except the lowest. This is because at scales short of the optimum scale the lowest cost for any output is to be had by "underusing" a plant with a capacity slightly larger than the output needed—by operating a certain plant at an output short of that for which its own average total costs are a minimum. With costs declining in response to increasing scale, the output for which the plant's average total costs are a minimum can be produced more cheaply with a slightly larger plant operated at slightly less than the larger plant's minimum-cost output. Past the optimum scale it is conversely most economical to "overuse" plant with a capacity slightly smaller than the needed output. (The capacity of a plant here refers to the output of the plant which gives it its own minimum-unit-cost rate of utilization.) For further discussion of the envelope curve, see Stigler, *op. cit.*, pp. 138-142.

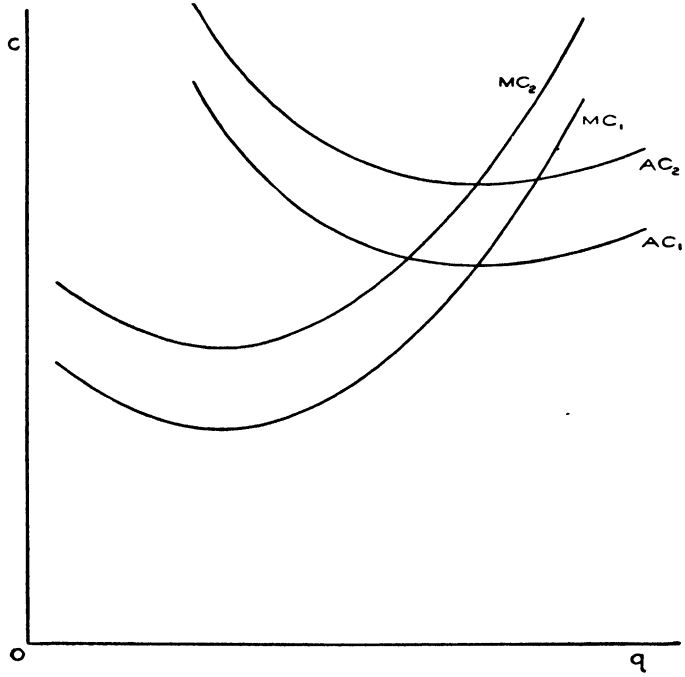

Figure 15

cause the firm to use different proportions of the various factors). Where wages and material prices respond to changes in the firm's output, of course, the effect of such responses must be reflected in movements along the cost curve. This is discussed further in Chapter 7.

There are obviously also other changes influencing costs which we have temporarily neglected in emphasizing the net relationship between cost and output. Two of these are changes in product and changes in the technique of production. In drawing our conventional cost curves we assume that the firm is producing a given definite product; the cost curves show us the relation of cost to output for that specific product. Variations in the design or quality of such product may well affect costs, but such effects are not shown in the cost curves, which specifically apply to one unchanging product. If the product is changed in quality or design, this must also be reflected in a shift to a new family

of cost curves, as from AC_1 to AC_2 in Figure 15. Both the position and the shape of the cost curves may, of course, be affected. The shift of costs in response to variation of product may be just as significant to the firm in deciding its policy as is the cost-output relation shown by the shape of any one curve.

The same may be said of changes in technique; the cost-output relation represented in any cost curve is linked with and reflects a specific technique of production. Changes in cost resulting from change in technique must be reflected in shifts between cost curves.

The advantage of showing separately in a cost curve the simple net relation of cost to output, and of representing any other changes in cost by shifts in cost curves, is that it enables us to consider the variables which cause costs to change one at a time, and thus to simplify the problem. It also enables us to describe cost variations on a simple two-dimensional diagram, rather than having to use multidimensional mathematical equations. It must be recognized, however, that the cost of production for a firm is jointly and simultaneously influenced by the rate of output, the type and quality of product, the type of technique, and the wage-price level at which it hires and buys. The firm in practice must take simultaneous account of all such variations, and of their effect on costs. It does not choose a rate of output *in vacuo*, or a technique *in vacuo*. It must make a joint decision on the technique to use, the quality and design of product to produce, and the rate of output to maintain, and it must consider the combined effect on cost of the various aspects of its total decision. This must be kept clearly in mind when we employ cost curves in analyzing price determination.

The student will also recognize that the arbitrary dichotomy of cost-output relations into a "short-run" and a "long-run" is a drastic simplification made for purposes of elementary analysis. The firm in practice, starting at any given date, considers many successively longer "short runs," each including the previous and extending a little further into future time, until it reaches one long enough to permit variation of all factors and thus to qualify as a "long run." Its decision-making process must in essence consider the relationship of a series of such overlapping time periods extending up to the "time horizon"

of its anticipations. When we speak of what the firm does in "the short run" and "the long run," we draw an extremely simplified picture of the decision-making process, albeit one which may shed much light on a more complicated reality.

It is further worth emphasis that the firm's cost curves, like the demand curve for its output, are operationally derived from *estimates* made by the firm. The cost-output relation which influences its decision is an *anticipated* relation, just as is its price-sales volume relation. Such estimates are probably drawn from past experience, but they are nevertheless estimates and subject to uncertainty. Where the uncertainty is great, the firm may have a rather sketchy knowledge of its strategic cost and demand curves, or may try to find ways of compensating for its lack of substantial knowledge of these determinants.

Finally, it should be emphasized that so far we have discussed only the relation of price to sales volume (the demand curve), and of production cost to output (the cost curve), on the assumption of many other things being given. We have also referred to the relation of product to production cost, of technique to production cost, and of money factor prices to this cost. But we have not discussed other relationships to which the firm must be sensitive—that of product to demand and of selling cost to demand for a product.

We have been concerned only with the costs to the firm of producing goods and making them available, for example, at the factory gate; we have not considered or included the costs of inducing people to buy them. Advertising, sales promotion, shipping charges borne by the seller, and allied expenses are selling costs. The distinction made is useful for two reasons. First, the importance of selling costs varies greatly from industry to industry and firm to firm; they may be negligible in some cases and important in others. Second, although there is a fairly clear relation between output and production cost for most firms (represented in the U shapes of average cost curves) there is no such predictable relation between output and selling cost. Sales promotion is not a stable technique which gives rise to predictable regularities in the relation of output to cost. In fact, selling costs are related to output only indirectly via their relation to the demand for the firm's product. It is therefore

appropriate that they should be given separate treatment, in the course of analyzing price determination. We will therefore turn at once to problems of price determination and discuss selling costs as those problems are analyzed.

SUPPLEMENTARY READINGS

JACOB VINER, "Cost Curves and Supply Curves," *Zeitschrift für Nationalökonomie*, III, 1932.

ALFRED MARSHALL, *Principles of Economics* (8th ed.), Books IV and V.

E. A. G. ROBINSON, *The Structure of Competitive Industry*, London: Nisbet, 1935.

CONFERENCE ON PRICE RESEARCH, *Cost Behavior and Price Policy*, New York: National Bureau of Economic Research, 1943.

KENNETH E. BOULDING, *Economic Analysis*, New York: Harper & Brothers, 1941, Chaps. 22-23.

CHAPTER 4

PRICE DETERMINATION IN PURE COMPETITION

Since the purpose of price analysis is to determine how a price system functions to allocate resources and distribute income, the nature of demands for and costs of products are not essentially important matters in themselves; a detailed knowledge of demands and costs, however much it is buttressed by mathematical refinement, has a distinctly limited usefulness. Demand and cost are important mainly because they are essential to the explanation of how the price system works.

Price formation begins with the individual firm, whose pricing decisions turn primarily upon the demand for its product and the cost of producing it. Let us now consider, in a simplified case, how such a firm makes decisions concerning price and output.

PRICE DETERMINATION FOR VARIOUS TIME INTERVALS

The most important problem concerns the function of the price system over time intervals long enough that the rate of production of goods can be varied and can thus become adjusted to the current levels of demand and of cost. We are thus concerned with price determination by the interaction of production costs and buyers' demands, an interaction which requires a sufficient passage of time for rates of output to change and for both

demands and costs to make themselves felt. We will not consider at any length price determination in *very* short periods, during which the rate of production cannot be varied, and for which, in effect, costs of production are less important in price determination than the size of stocks or inventories on hand.

We have already mentioned two sorts of time intervals for which cost variations may be analyzed: (1) the long period, during which the firm may vary output through free variation in the quantities of all productive factors employed, and (2) the short period, during which output can be varied through variation in quantities employed of only a part of the factors of production, certain "plant factors" being fixed in quantity. In either of these periods costs of production interact with demand to determine price and output. For long-period planning, expected long-run average and marginal costs interact with the anticipated long-run average level of demand to determine a long-run central tendency for price and output. During short periods, short-run average costs interact with the currently prevailing (and changing) levels of demand to determine specific price tendencies for these shorter periods. A series of short-period prices, following a fluctuating demand, may have as their central tendency the long-run price just mentioned.

There is, however, an even shorter interval, which we will simply call a "*very short*" *period*, short enough that for its duration output cannot be varied at all. For such a period, price depends mainly on immediate demand and on the amount of stocks on hand. If demand fluctuates during such periods, price movements will depend on the amount of inventories on hand and on the rate at which the owners of such inventories are willing to dispose of them. Such very-short-period prices may tend to fluctuate on either side of short-run prices.

The theory of very-short-period pricing will not be expounded in much detail. In a very short period there will be for any good a given market demand, showing the amounts buyers are ready to take at various prices. There will also be a certain stock of the good available or on hand over a time interval during which existing stocks of finished output cannot be supplemented. Dealers in the good may decide to put the entire stock on the market during this very short period, in

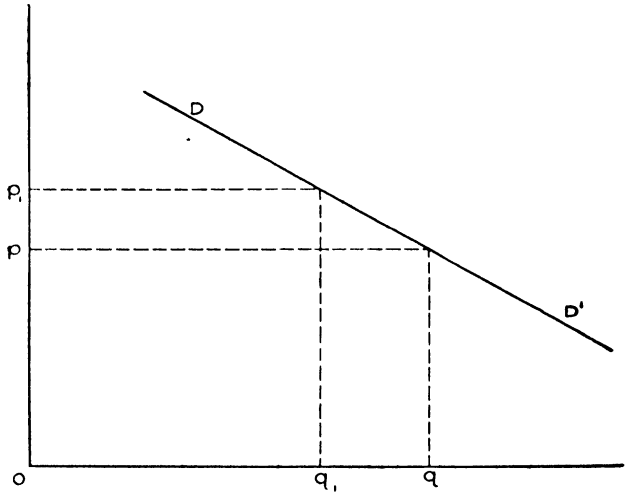

Figure 16

which case price will find a corresponding level on the demand curve, or they may decide to withhold part of it, in which case the price will be higher. Thus if in Figure 16 the demand for a good is *DD'* for a very short period, and the stock on hand is *oq*, offering of the entire stock will result in a price of *op*. But if sellers offer only oq_1, withholding the amount q_1q, the price will be op_1. Various suppositions concerning the tendency of sellers to withhold stocks may be developed. By and large, however, their tendency to withhold will be governed by the prospective *short-run* price, which is expected to prevail as additional output is forthcoming and which will rule as existing inventories are replaced. Very-short-period prices are thus tied fairly closely to short-run prices and to costs. We will dispense with further discussion of special theories of very-short-run prices and turn directly to price determination as related to costs of production.

THE SIGNIFICANCE OF PURE COMPETITION

The process of such price formation, and indeed the prices which result from it, differ significantly according to the sort of

market structure within which individual sellers operate. The characteristics of this market structure influence the nature of the demand for the individual seller's product, with the result that price formation and price results are different for different types of industries. This was suggested at the end of Chapter 2 in our discussion of individual sellers' demands. We must therefore consider separately the nature of price determination in each of several types of industry—initially in pure competition, monopoly, differentiated and pure oligopoly and monopolistic competition.

We begin with industries in pure competition—where there are many small sellers, all of whom produce identical products. Several markets for agricultural crops, especially the grains, reasonably approximate purely competitive conditions. That is, a given crop is produced by several thousand farmers, no one of whom controls a significant proportion of the total output. Their various outputs, moreover, being reduced to standard specifications or grades, are viewed as identical by all buyers. In the field of industrial production, purely competitive market structures are unusual, because ordinarily the sellers are few and their products are often differentiated. Two industrial markets which give fair approximations to pure competition, however, are that for cotton gray goods and that for bituminous coal. Cotton gray goods are produced by about six hundred small firms, without any very significant differentiation among their outputs. Soft coal is produced by over a thousand small mining companies, and the output is substantially undifferentiated. But for the American economy as a whole, the purely competitive market or any reasonable approximation to it is an exceptional case.

Some explanation is therefore required for treating this rather special case before turning to the more common market situation where sellers are few and products significantly differentiated. Pure competition is considered first for two reasons. First, the analysis of pricing in pure competition is relatively simple and uncomplicated, and thus serves as a useful introduction to more complicated phases of price analysis. The range of pricing and allied decisions which must be made by a seller in pure competition is much narrower than that facing sellers in monopo-

listic and oligopolistic markets. There is thus less for analysis to take account of, and certain fundamentals common to all markets may be made to stand out more clearly. Second, the price behavior and results which emerge from purely competitive markets serve as a convenient measuring rod or standard for appraising the price results in other (and more common) sorts of markets. The price results associated theoretically with pure competition are often held to possess certain *normative* properties, or to represent a sort of ideal in capitalist pricing behavior. In any event, many of the traditional justifications for a *laissez-faire* economy, which argue that a capitalist economy is through its price system automatically self-regulating toward ideal results, refer explicitly or implicitly to a world of purely competitive markets, or at any rate of markets not *significantly* different in their operation from markets in pure competition. This is true of Adam Smith (*The Wealth of Nations*, 1776) and of Friedrich Hayek (*The Road to Serfdom*, 1944).

Without prejudging the normative merits of pure competition, we recognize here that the analysis of purely competitive markets may provide tentative standards by which other more common sorts of behavior may be measured. Also, of course, there are enough purely competitive markets in the modern economy to make investigations of this type practically important. We will therefore consider at once the operation of the firm, the industry, and the economy when governed by pure competition.

SHORT-RUN PRICING BY THE FIRM IN PURE COMPETITION

The basic characteristics of a market in pure competition are (1) that the various sellers in the market produce a single identical product, and (2) that they are so many in number and so small that no one of them can perceptibly influence the price of this product. From the large number and relative smallness of these sellers there may also ordinarily be inferred a third market characteristic. It is quite easy for additional sellers to enter the market if they so desire; there is *ease of entry*.[1] In such a

[1] The full conditions for "easy" or "free" entry may be defined as follows: (1) the investment necessary to establish a firm is small enough that many

market, how may price and output be expected to behave, and how may the force of competition be expected to govern productive activity?

There are in general two ways of seeking an answer to this question. One is to find some markets with a purely competitive structure, observe what happens in them, and then make some generalizations concerning their behavior. Another is to postulate the relevant conditions which control the operation of enterprise in such a market, and to work out deductively the sort of behavior which logically should emerge from the postulated circumstances. The latter method is that of conventional price analysis. It is certainly not a substitute for inductive investigation, which should always supplement and test the conclusions of the more abstract theory. But it does offer by far the most facile and accessible means of developing a general idea of what the significance of any market situation is. We will therefore turn immediately to the more or less abstract analysis of purely competitive pricing, and will comment a bit later on the inductive examination of the process.

Our first inquiry must concern how the individual firm in a purely competitive industry decides upon its price and its output. Suppose, for example, there are 600 small sellers of gray cotton yarn, and consider the price-output decisions of one of them—first for a short period. Here we have a firm with a given product (a standardized grade of yarn) and a given fixed plant, which for the course of the short period it cannot vary in size.

potential entrants can obtain sufficient funds to establish a firm, (2) the increment to industry output resulting from the entry of one additional firm is so small as to have no perceptible effect on industry price, and thus the potential entrant is not deterred by fear of changing the existing price situation; (3) all potential entrants have free access to all resources or factors needed for production, at competitive market prices, since there is no monopolization of resource ownership or control by established firms; (4) there are no other artificial impediments to entry; and (5) new entrants can produce outputs identical to those of established firms. Where all these conditions are observed, we have the ultimate in ease of entry, or purely competitive free entry. In markets outside the purely competitive category, however, there may be *relatively* easy entry, where some but not all of these conditions are observed. "Easy" entry in oligopoly, for example, might observe all conditions except the second, and possibly the first and fifth.

As a small buyer of factors, it will also be faced with given money prices for the factors of production which it purchases and which enter into its costs of production. How does it decide how much to produce and what price to charge for the next short-run (*e.g.*, six-month) interval?

Presumably the firm will calculate (1) the prospective selling price it can receive for its goods and any variations of this price with variations in its own output and sales volume, and (2) the prospective cost of producing goods for sale, and any variation of this cost with variations in its rate of output. In short, it will refer to the demand curve for its own output and to its short-run cost curves.

The demand curve for the output of a seller in pure competition is simply a horizontal straight line at the level of going or anticipated market price. That is, he can sell any output he can reasonably produce at the market price without causing the price to change perceptibly, and can sell nothing above that price. The seller in pure competition therefore has no price decision to make. He will simply accept or estimate market price and choose the rate of output which, in the light of the relationship of the production costs to output, seems to him most profitable.

The nature of his problem may be illustrated by considering the going price of his product in conjunction with the short-run cost curves for his given plant in a single diagram, as in Figure 17. Here *dd'* is the demand curve for the firm's own output—in effect a "price line" showing the going market price at which it can sell any output. The various cost curves show how its average total cost, average variable cost, average fixed cost, and marginal cost of production, vary with varying output, given the prevailing factor prices. The usual motive of a business firm will be *to choose that output which will maximize its profit for the period under consideration.* The firm supposedly seeks the output which, for the short period at hand, will allow it to make the largest possible aggregate profit or smallest possible aggregate loss.

In Figure 17 it is apparent that the firm can make some sort of profit at any output between q_1 and q_2 (at which limits price

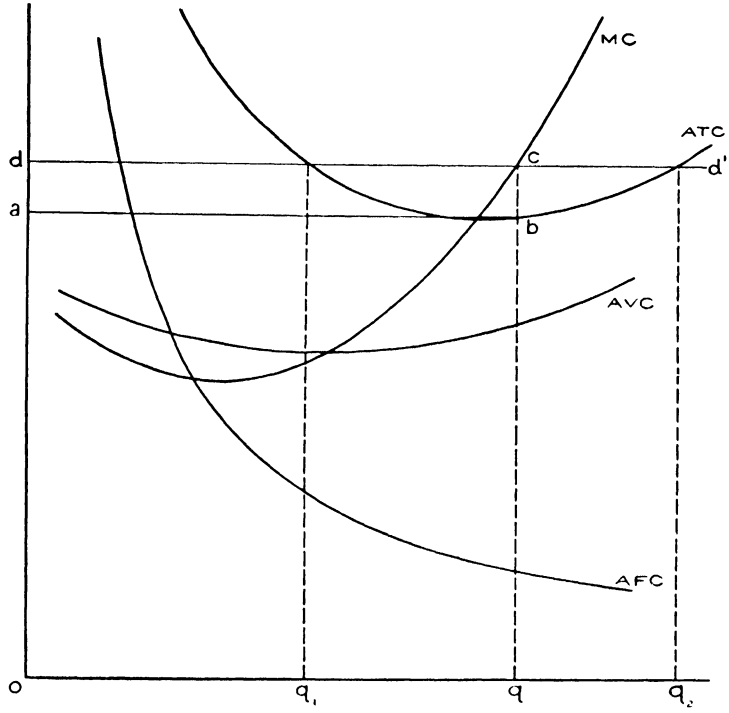

Figure 17

is equal to average total cost) though at outputs smaller than q_1 or larger than q_2 it would make a net loss. To choose the output (between q_1 and q_2) which will maximize its profit, the firm will refer directly to its *marginal cost of production* and the *marginal receipts from its sales.*

The *marginal cost* of production, shown by the line *MC*, is the addition to cost for any unit addition to output. The *marginal receipts* is now defined correspondingly as *the addition to total revenue for any unit addition to output* (and sales). In pure competition, where price remains invariant as the individual seller increases his output, his marginal or additional receipts from any unit added to output are the same as the price of that unit of output. Thus if a firm's demand schedule shows the following purely competitive relation—

p	q	Total receipts $p.q$	Marginal receipts
10	12	120	—
10	13	130	10
10	14	140	10

it is apparent that marginal receipts are always the same as price. The firm's demand curve, *dd'*, is therefore also its *marginal receipts curve*, showing additions to total revenue per unit of additions to output.

To find the most profitable output, the firm will logically add to its output as long as the marginal cost of additional output is less than the marginal receipts of that output. But it will not extend output when the marginal cost of additional output becomes greater than the marginal receipts. It is then evident that the firm will maximize its profits at the output for which marginal cost just equals marginal receipts. In pure competition, where the marginal receipts earned by any unit of output are the same as its price, the firm will evidently extend output exactly to the point where marginal cost equals price. This is the output *oq* in Figure 17.

At this output the price is measured by the distance *qc*, the total cost per unit by the distance *qb*, and the profit per unit of output by the distance *bc*. The *aggregate profit* is the rectangle *abcd* (the profit per unit, *bc*, multiplied by the quantity of output, *ab*). This is necessarily the largest profit which can be earned during the short period in question. It is suggested that the student reconsider the preceding four paragraphs until he understands the argument thoroughly.

The general rule which the firm follows to maximize its short-run profit is to select the output at which marginal cost is equal to marginal receipts, which in the case of pure competition also means that marginal cost is equal to price. If the market price at which the firm can sell changes during a short period, the firm will change its output sufficiently to bring marginal cost into equality with the new price.

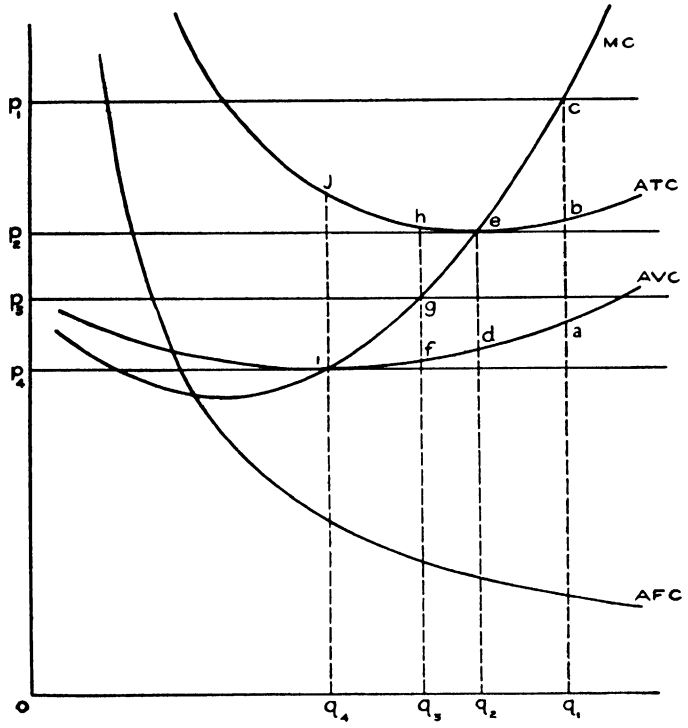

Figure 18

In Figure 18 we suppose that a firm, with the same costs shown in Figure 17, is faced successively with the prices p_1, p_2, p_3, and p_4. As the market price changes from one level to another, the competitive firm's demand curve, a horizontal line, simply shifts to a new level. It is evident from our preceding argument that at the price p_1 the firm will produce q_1; at p_2 it will produce q_2; at p_3 the quantity q_3; at p_4 the quantity q_4. That is, it will adjust its output so as to *equate* marginal cost to each new price. In this way it will earn the largest profit or incur the smallest loss obtainable at any particular price.

It is not necessary that the firm make a profit in excess of full costs, or that it break even, to induce it to remain in business for the short period. (The average total costs of the firm, represented in the *ATC* curve, are defined as inclusive of a normal

return on capital and of wages of management—or, in effect, of a normal or necessary profit.) Whether the firm makes more or less than costs including this normal profit in a given short period simply depends upon where market price happens to lie. If market price is high, as at p_1, the firm makes an excess profit at its most profitable output, in the amount *bc* per unit. If price falls just at p_2, the firm can just recover full costs (average total costs equaling price), but can make no excess profit. If price is p_3, the firm makes a net loss in the amount *gh* per unit. If price is still lower at p_4, it makes a larger net loss of *ij* per unit. Any of those results might ensue in the short period. Higher prices ordinarily characterize prosperous and lower prices depression periods.

It is easily evident that the firm should be willing to produce at the prices p_1 or p_2, where it either makes an excess profit or recovers full costs. But it may at first appear surprising that the firm should continue to produce at a net loss, at prices p_3 or p_4. The logic of such a procedure is evident, however, if we recall that short-period costs are partly fixed and partly variable, and that the fixed costs would be incurred at zero output. If the firm refuses to produce at all, it will incur a net loss equal to its total fixed cost. It should therefore be willing to accept any price which will enable it to reduce this loss or, in other words, any price in excess of the average variable (or out-of-pocket) cost of production. If it does this, it will find that production at a net loss per unit nevertheless enables it to minimize its total loss for the short period, by recovering some proportion of the otherwise lost fixed cost.

Thus in Figure 18 the firm is willing at the price p_3 to produce the output q_3 (making a net loss per unit of *gh*), because this enables it to make a return per unit *above average variable costs* in the amount *fg*. This return, though not enough to defray all fixed costs, at least recovers a part of them and thus makes production worthwhile. We may therefore generalize that the firm will produce in the short run (at an output where marginal cost equals price) as long as price exceeds the average variable cost of output. Thus the minimum price which would induce the firm to produce in Figure 18 is p_4, where price just

equals average variable cost. At lower prices the firm would presumably shut down.[2]

The excess above variable costs which the firm is able to earn in a short period is often referred to as the *quasi rent* earned by the fixed plant in that period. Thus in Figure 18 at the price p_1 the quasi rent per unit of output is *ac*; at p_2 it is *de*; at p_3 it is *fg*; and at p_4 it is zero. The rule is that the firm will produce in the short period as long as it can earn any positive quasi rent.

It should also be evident, since the competitive firm always adjusts its output in accordance with the variation in its marginal cost of production, that the firm's rising short-run marginal cost curve may be regarded also as its *short-run supply curve*—that is, it is the curve which indicates the amount of output forthcoming at any particular price. This is precisely true for the range of the firm's marginal cost curve wherever it lies above average variable cost.

So far we have described the way in which in the short run a firm in pure competition should adjust its output to the going market price. Such a firm has no control over price, and therefore no "price policy." It necessarily accepts the going or estimated market price as given and beyond its own control, and devotes itself to selecting that output which offers the largest short-run quasi rent and profit. This output is necessarily defined as that for which the firm's short-run marginal cost of production equals price. The firm will produce so long as, at this output, price is greater than average variable cost.

THE SHORT-RUN PRICE FOR THE INDUSTRY IN PURE COMPETITION

Although our analysis thus far suggests how a single firm should respond to a given market price, it does not yet tell us

[2] How far the firm will see price drop below its full cost before shutting down depends upon the fraction of its full costs which it considers fixed for such a period. This proportion is partly a technical matter, depending upon the relation of equipment costs to labor and material costs. But it also depends upon the length of the short period for which the firm calculates. The longer a short period it takes into account, the fewer of its costs will appear fixed, or, in effect, the less it will be willing to lose now in order to be in operation for a succeeding period.

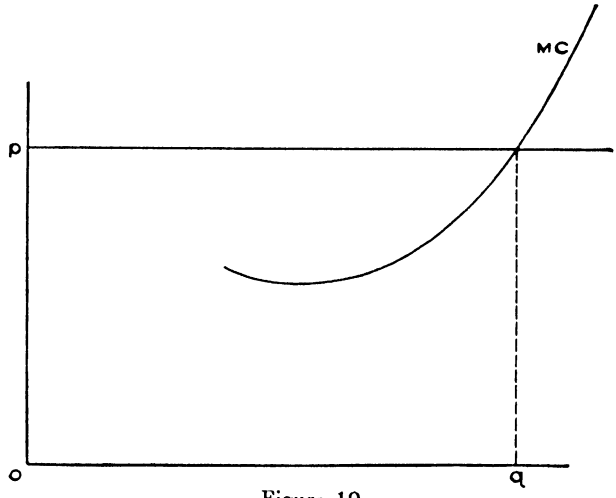

Figure 19

how this market price is arrived at. The next question therefore is this: How in a purely competitive market do the aggregated reactions of individual firms determine the industry price? We consider this question first for the short run, during which *the number of firms* and *the sizes of their respective plants* cannot be changed.

A purely competitive market is made up of a large number (let us say 700) firms producing a single product. On a certain date the market price of their good—for example, gray cotton yarn—is *op* per unit. At this price each firm produces some output *oq*, at which its own short-run marginal cost equals this price, as in Figure 19. What happens in the market as a whole? For the entire market an *aggregate supply* equal to the sum of all the individual *oq*'s of the several hundred firms is produced. If there are 700 firms, the aggregate supply may be defined as:

$$Q = (q_1 + q_2 + q_3 + \cdots q_{700}),$$

or, in effect, the sum of the individual outputs chosen by all firms.

This aggregate supply, together with the *industry demand* for the product, should establish a market price. The industry demand for gray cotton yarn might appear as *DD'* in Figure

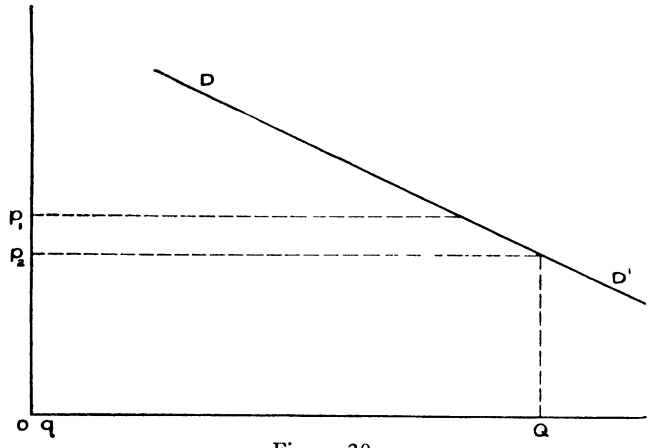

Figure 20*a*

20*a*, showing the amounts of yarn which buyers as a whole are prepared to take at each market price. The price scale in Figure 20*a* is the same as that shown for the individual seller's demand line. The quantity scale runs in much larger quantities. Thus the output *oq* of one seller is represented by a very small distance, whereas the combined output of 700 sellers is the much longer distance *OQ*.

In the example in Figure 20*a*, we began by supposing a provisional price op_1 (similar to *op* in Figure 19) for the good. At this price sellers react and produce a combined output *OQ*. Now the process of market price determination is under way. The output *OQ* will sell at some specific price, but not necessarily at the provisional price op_1 which brought it forth. In the example above, it will sell at a lower price, op_2 and market price will therefore fall to this level. But with this price drop each seller will produce a little less (reducing output to a point where marginal cost again equals price). The aggregate output will therefore be reduced to somewhat below *OQ*, and price will rise above op_2. By a process of successive adjustment the market will arrive at some price op_e, intermediate between op_2 and op_1, *at which the aggregate amount buyers are willing to take just equals the aggregate amount sellers are willing to supply.*

Thus it is clear that the aggregate of adjustments of supply

undertaken by many small firms, interacting with the market demand, determines a short-run "equilibrium" price at which industry demand and supply are in balance. This explanation is accurate enough as far as it goes, but it is unnecessarily cumbersome. The determination of short-run market price in pure competition may be characterized more concisely as follows. The demand conditions affecting price are evidently represented in the industry demand schedule for the good, *DD'*, showing the amount which all buyers will take at each price. The short-run supply conditions for the market as a whole may be represented in a corresponding *industry supply curve SS'*, which shows the amounts of goods all sellers will offer at each price. *This short-run industry supply curve is evidently the aggregate of the short-run supply curves of all the individual firms in the market, or, in effect, the sum of the short-run marginal cost curves of these firms.* The supply curve of the individual firm in pure competition is that range of its rising marginal cost curve which lies above its intersection with the firm's average-variable-cost curve, since this segment of marginal-cost curve shows the amount of output the firm will produce at each corresponding price. To get the industry supply curve we simply make a "horizontal" addition of all such individual marginal-cost curves, and arrive at a summation which shows the aggregate amount all firms will furnish at each possible price.

If we now, as in Figure 20*b*, place this industry supply curve, *SS'*, in juxtaposition with the market demand curve, *DD'*, the intersection of the two curves shows the equilibrium market price. In the short period the equilibrium price will be op_e, at which the amount supplied by all sellers just equals the amount demanded by all buyers. Departure from this price in either direction will set in course adjustments which tend to return price to the equilibrium level.

(The preceding construction of the short-run industry supply curve assumes that variable factor prices paid by firms in the industry do not change in response to a change in industry output. The supply curve thus is a simple summation of firms' marginal cost curves, each drawn on the assumption of given factor prices, and its slope reflects only movement along such curves. If, however, factor prices change in response to changes in industry

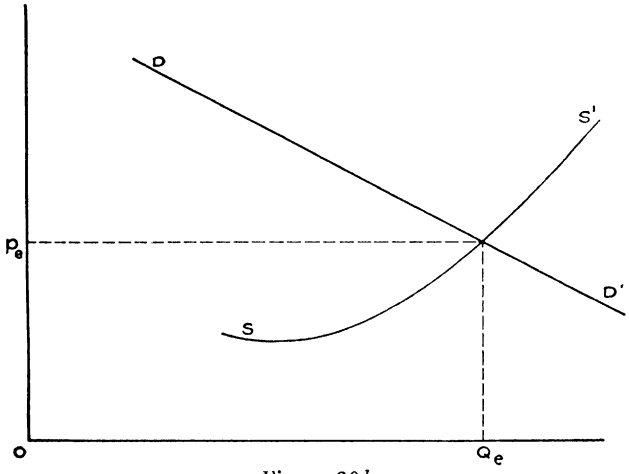

Figure 20*b*

output—though not to changes in any firm's output—thus causing the cost curves of all firms to shift as industry output varies, then the industry supply curve must be drawn to reflect such induced shifts in firms' marginal cost curves as well as movements along them.)

The equilibrium has two essential properties. First, price is such that the aggregate amount sellers are willing to produce just equals the amount buyers are willing to take. The price thus tends to be maintained rather than to be departed from, so long as demand and supply conditions are unchanged. Second, each seller is in individual short-run equilibrium, producing just such an amount that his marginal cost equals market price. This is of course essential if the price is an equilibrium price.

It is thus clear in a purely competitive market, although each seller regards price as outside his control, and simply adjusts his output to whatever the going price may be, that the combined actions of many such sellers unequivocally determine a definite market price at which the individual profit-seeking adjustments of all sellers can be maintained consistently with the aggregate desires of buyers. We thus see, first in the instance of short-run adjustments, that a market in pure competition is "automatically regulated" or "self-regulating" to a certain end. "An invisible

hand," as Adam Smith described it, harnesses the essentially selfish adjustments of each of many sellers to produce a price result that none of them has planned.

What is the character of this unplanned price result? For the short run, it has relatively few properties. First, the equilibrium price is equal to the marginal cost of production of all sellers. Second, the short-run price may be equal to, greater than, or less than the average total cost of any or all producers. As long as all plant capacity and the number of firms are fixed, a short-run equilibrium can be maintained even though net losses are general, or though highly attractive excess profits are generally being earned in the industry. Adjustments to correct for such conditions can occur only in the longer run. All that is required in the short run is that price shall be at least as high as the average variable costs of production of at least some sellers in the industry. In the short run, price bears no necessary correspondence to average total cost of production.

These are the principal evident properties of purely competitive price in any given short-run situation. Certain other significant properties are observed, however, if we consider short-run adjustments to changes in demand or cost. As the economy moves through time, the level of demand for goods generally or for any particular good ordinarily fluctuates with changes in the volume of consumer purchasing power. Similarly, the level of industry costs fluctuate with changes in wage rates and material prices which are independent of the industry's output. Such fluctuations are reflected in systematic *shifts* in industry demand and supply curves, as shown in Figure 21. Thus an increase in demand with the advent of more prosperous times would be reflected by a shift of the industry demand curve for a good from DD' to D_1D_1'. A rise in the level of costs independently of industry output would be registered in a shift of the industry supply (combined marginal cost) curve from SS' to S_1S_1'. When such shifts occur in a purely competitive industry, no seller can control price, and as a result there is an immediate and full adjustment of price and output to changes in demand or cost. Thus if demand is DD' and supply SS', equilibrium price is op and quantity oq. But a shift of demand to D_1D_1' (supply remaining the same at SS') will automatically raise price to op_1 and

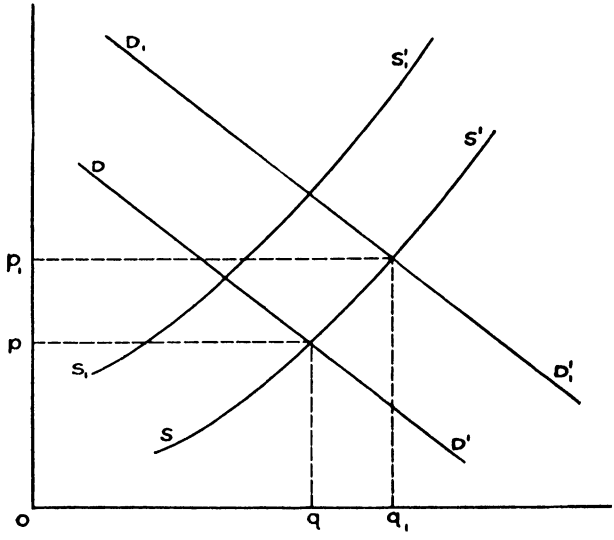

Figure 21

quantity to oq_1. Other shifts in supply or demand will bring similarly immediate responses.

An important property of pure competition is thus that price is automatically responsive to changes in demand or cost, and tends to be very *flexible* over time if there are changes in these price-determining variables. The purely competitive price thus tends to perform an active regulatory function in adjusting output to changes in the surrounding economic situation. So far as it tends to rise and fall readily in response to controlling changes, moreover, it tends to dampen fluctuations in output, which thus can be more stable over time than it would be if price were more rigid.[3]

The preceding discussion has characterized the determination of price and output in pure competition for the short run—that is, for periods of six months to one or two years during which the amount of fixed plant available in such an industry is relatively inflexible. For such a period, price and output observe

[3] For a general discussion of price flexibility see Saul Nelson and W. G. Keim, *Price Behavior and Business Policy*, Temporary National Economic Committee, Monograph No. 1 (Washington, 1941), Chap. 2.

certain systematic tendencies. In particular, they are readily adjusted to the marginal costs of producing with the given fixed plant. Otherwise, however, short-period price behavior follows no certain pattern. The volume of output, the size of profits, and the relation of attained to optimum (lowest-unit-cost) output varies with the relation of the amount of productive capacity to the current level of buyers' demand. But over longer periods, during which the size of plants and firms and the number of firms in the industry can be freely increased or decreased, we should expect that the force of competition would effect some more certain average relation of capacity to demand and place some limits on the average rate of profit or loss incurred by producing firms.

THE LONG-RUN PRICE FOR AN INDUSTRY IN PURE COMPETITION

For a long period as for a short one, the firm in pure competition is concerned with adjusting its rate of output to the going or anticipated level of market price. But in the long run it may vary its output not only by varying the rate of use of its existing plant, but also by simultaneously varying the size of this plant. The relationship of cost to output in which it is interested is therefore described by its long-run cost curve, already discussed in Chapter 3. The firm will presumably set its long-run average and marginal costs against the expected long-run level of market price and choose that combination of plant size and rate of use of plant which it believes will afford it the greatest profit.

The firm's adjustment is seen in Figure 22. Its demand curve, as always in pure competition, is a horizontal line, at the expected long-run average market price. It considers its long-run average and marginal costs in conjunction with this price. Following the logic already developed, the firm will set its scale of operations at such a level that long-run marginal cost is equal to price (since this allows the maximum long-run profit) so long as price is equal to or greater than its minimum average cost for the long run. That is, at the price p_2 it will build to a scale to produce the output q_2—to a larger than optimum or

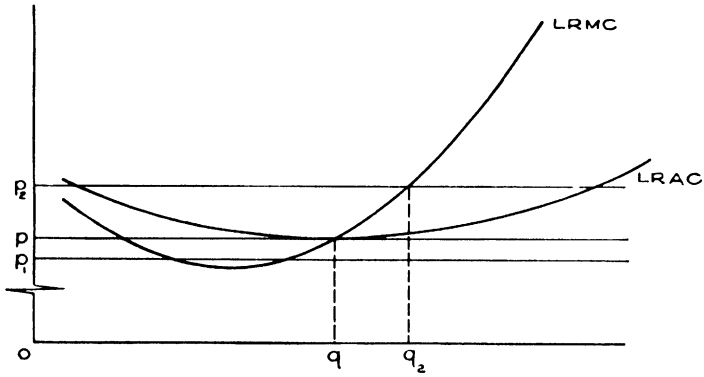

Figure 22

lowest-cost scale. At the price p, it will build to a scale to produce the output q—to optimum scale. At any lower price p_1, this firm will not produce at all; it will refuse to build any new plant or will begin to liquidate its existing plant.

So much for the reaction of the single firm to any anticipated long-run price, which is precisely the same order as its adjustment to a short-run price. The determination of the long-run price for the industry, however, does involve a different element and a different result. The industry price will in effect be determined jointly by the aggregate output adjustments of an initial number of firms in the industry *and by the entry or exit of firms in response to excess profits or net losses.* Suppose initially, as in Figure 23, that the long-run industry demand is DD', that ΣMC represents the summation of the long-run marginal cost curves of the existing firms in the industry, and that AC is the level of the minimum average cost (including normal profit) of every firm in the industry. (The reason for considering all firms equal in this respect will be discussed below.) In this event, the combined industry output will be OQ, the price, op, will be equal to AC, and all firms will be making just a normal profit at minimum cost. Each firm will be in a position indicated by the relation of p to $LRAC$ in Figure 22. Since none will wish to leave, and since additional entry would cut price below minimum average costs, this is evidently a long-run equilibrium output for the industry.

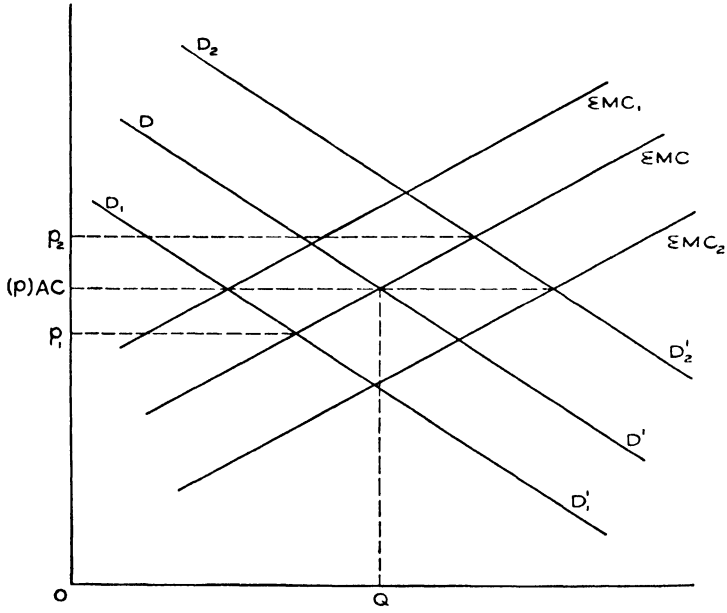

Figure 23

Such an equilibrium would not ordinarily occur without adjustment, however, but would have to be approached by entry or exit. Suppose, alternatively, the number of firms and the conditions of cost being the same, that the demand for the industry is D_1D_1' (Fig. 23). The initial number of firms (with combined marginal costs ΣMC) would now be too large—that is, the price op_1, which would result if all equated marginal cost to price, would fall below minimum average costs, and firms would begin to leave the industry. Each firm would be in a position indicated by the relation of p_1 to $LRAC$ in Figure 22. As this occurs, supply will be restricted, the ΣMC curve will shift leftward, and price will rise. Where it reaches the position ΣMC_1, price will equal op, equal to minimum average costs AC, and the industry will be in long-run equilibrium to the demand D_1D_1'. (Price will in effect be p from Figure 22.) In short, exit of firms takes place until total supply is such that price equals the minimum average cost of productions of all firms, and here exit stops.

Suppose, alternatively, that with the same minimum average costs, and with the initial aggregate marginal cost ΣMC, the industry demand is D_2D_2'. Now with the initial number of firms, price would be op_2, and above the average costs of all firms, in the position shown by the relation of p_2 to $LRAC$ in Figure 22. These firms will therefore be content to remain and make excess profits, but additional firms will wish to enter the industry to get some excess or supernormal profits also. As they do so, the aggregate supply (ΣMC) will shift rightward and price will fall. When it reaches ΣMC_2, price will be op, equal to minimum average costs, AC (the minimum $LRAC$ from Figure 22), and the industry will be in equilibrium to the demand D_2D_2'. The entry of firms takes place until total supply is such that price equals minimum average costs.

Taking the preceding argument as a whole, it is evident that in the long run, and with unimpeded entry and exit of new firms, the scale and number of firms will adjust together to such a level that price equals the minimum average costs of production of all firms. This is the result of free and unrestricted pure competition. In arriving at this conclusion, however, we of course made certain simplifying assumptions—namely, that all firms in an industry have the same minimum average cost ($LRAC$) at any given industry output and that this minimum cost does not shift up or down as industry output is extended or restricted by entry to or exit from the industry. Where these conditions both hold, *the long-run supply curve for the industry* is in effect a horizontal line at the level of this minimum average cost—for example, the line extended from AC in Figure 23. That is, under free entry and exit firms are in the long run continually forced to produce at prices equal to their minimum average costs, and output will be extended or retracted at this uniform minimum cost-price level simply by adding to or subtracting from the number of firms until output is such that price equals this cost. Hence the locus of successive minima of the long-run average costs of the member firms (AC in Figure 23) is the long-run supply curve for the purely competitive industry. Where the demand curve intersects this supply curve we find the long-run price and output for the industry. Although each firm adjusts so as to produce where long-run mar-

ginal cost equals price, the pressure of entry and exit is such that it succeeds in producing only where minimum average cost is equal to price. Since this minimum average cost is not altered as industry output is changed by entry or exit of firms, the industry has a perfectly elastic supply curve at its level.

Let us now reconsider the simplifying assumptions concerning uniformity of minimum costs as among firms at any given industry output and the constancy of the minimum cost level in the face of entry or exit. The proposition that all firms have identical minimum costs at any given industry output in the long run stems from the definition of costs. It is initially assumed for this argument that all firms can and do use the one most economical technique of production (this may be considered a corollary of free entry) and thus that their costs do not differ because of differences in technique. Given this, all firms must have identical minimum average costs in equilibrium so long as the units of any factor used by one firm have the same efficiency as those used by others, and so long as each firm pays or counts as the cost of any factor the same given market price that other firms do. If they use factors of differing efficiency, their minimum average costs will be identical if each factor unit is paid or imputed a market value representing the full earning power corresponding to its particular efficiency. An *apparent* difference in costs among firms may result from the fact that some firms in calculating costs impute or pay to factors of given quality other than their market prices, or from the fact that they do not pay or impute to factors of superior efficiency the full market value of their superior efficiency. When costs are defined to include a market value imputed to each factor unit which represents its full earning power (under conditions of competitive free entry)—as they should be—all firms in a competitive industry will have identical minimum average costs, which will equal industry price in long-run equilibrium. Costs will of course not include the earning power of impediments to entry, since these are assumed not to exist.[4]

[4] Cf. Stigler, *op. cit.*, pp. 159-166, and Boulding, *op. cit.*, Chap. 21, for further analysis of this somewhat complicated issue, which is treated only in elementary and simplified fashion here.

The second assumption was that the minimum average costs of firms do not shift up or down as firms enter or leave the industry—that as firms are added, they will not have higher minimum costs than those already in, and that as firms leave, this will not result in the remainder having lower costs. Investigation of this assumption leads us to consider the relationship of expansion or contraction in one industry to the prices of productive factors in the economy as a whole. If all productive factors are available to a given industry in perfectly elastic supply—if any of them may be obtained in larger or smaller quantity without influencing their money prices—then output can be expanded or contracted by entry or exit of firms without influencing the money costs of any firm or the proportions in which various factors are employed. This assumption of a perfectly elastic supply of all factors is thus the direct corollary of the assumption of a perfectly elastic long-run supply curve for the output of a purely competitive industry. If, however, some or all employed factors are in limited supply to the industry, so that their prices increase as entry takes place and more are used, and that their prices fall as exit takes place and less are used, then the money costs of all firms will shift up or down (*LRAC* for every firm will shift up or down) as entry or exit takes place. As a result, entry to eliminate excess profits will result in rising industry cost as well as falling price and increasing output, and exit to eliminate losses will result in falling costs as well as rising price and declining output. The long-run equilibrium for the industry will still be characterized by all firms producing with minimum *LRAC* equal to price. But the long-run industry supply curve (the locus of minimum *LRAC*'s at various outputs) will have an upward slope instead of being horizontal, and industry price in the long-run must be expected to rise if output is expanded.

What is the expectation of the character of this long-run supply for actual industries? For a single small industry buying nonspecialized productive factors in large markets, expansion or contraction of its requirements will probably not influence factor prices much, and industry supply should be very elastic. For industries using a factor in approximately fixed supply, like land

of a certain quality, expansion of output will raise the price of that factor (as well as involving changed factor proportions) and supply price will increase as output is increased. As we consider larger industries, or groups of industries together, it is apparent that general expansion of output, drawing on a supply of all factors which for the economy as a whole is definitely limited, may result in rising factor prices and costs unless there is a reserve of unemployed factors to draw upon. Any general expansion of output, in an economy where unemployed factors require higher than prevailing wages to induce them to work, will result in a general upward shift of costs and prices.[5]

Returning to the equilibrium of the single industry in pure competition, it is clear that several specific sorts of results tend to emerge from such an industry in the long run. We have already shown that in the short period the following tend to be true:

1. Firm output is always so adjusted that short-run marginal costs are equal to price.
2. Industry price and output tend to be automatically adjusted in accordance with the relationship of the short-run industry supply curve to demand.

In the long run, it is apparent that the following should also tend to be the average results:

3. Industry output is so adjusted that price equals the average cost of industry output, which is also (with one exception) a measure of the *socially relevant* marginal cost of an increment to industry output. (We will refer to this idea of socially relevant marginal cost below on p. 127 when the concept is put to use.)
4. Industry price tends to be at the level of the minimum average cost of production of all firms. The firms of the industry thus earn no excess profits, but only the normal

[5] The causes of slope in the long-run industry supply curve have hardly been exhausted here. For a more extended discussion, see Viner, "Cost Curves and Supply Curves," *Zeitschrift für Nationalökonomie*, III, 1932. See also Joan Robinson, *Economics of Imperfect Competition* (London, 1934), Macmillan & Co., Ltd., Ch. 10, for a further analysis of industry costs.

profits—wages of management, interest on investment, etc.—which are included in costs.

5. Since all firms are at their minimum average costs of production, competition has forced them to operate at their *optimum scales* and at their *optimum rates of utilization.* Minimum unit costs are thus secured.
6. There are no selling costs to be covered by price, since these are ruled out by the complete lack of product differentiation.

NORMATIVE PROPERTIES OF INDUSTRY BEHAVIOR IN PURE COMPETITION

If any sort of market is automatically regulated by forces of demand and supply, over which individual sellers have no control, it is a market in pure competition. The automatic regulation to which such a market is subject seems advantageous to the whole economy in its apparent effect on the efficiency of production and the distribution of output. Firms in the long run tend to be forced to build plants of optimum size and to use them at the optimum rate. This means that any output which the industry produces will be produced at the minimum attainable aggregate cost for that output. The freedom of entry, together with the multiplicity of firms, forces this. This minimum attainable aggregate cost for any industry output corresponds in pure competition to the optimum-scale costs of all firms, since industry output is adjustable in the long run by changing the number of firms rather than by varying the scale at which they operate.

Because price is forced to the level of average cost, moreover, there is a systematic limit on the profits of enterprise, and therefore on the distribution of the income of the industry between ownership interests and employed factors such as labor. The owners of firms in such industries tend to realize only a normal interest return on investment plus managerial wages plus any reward of superior efficiency, while other factors of production tend to get the balance. In particular, any excessive profits from artificial limitation of supply are effectively ruled out.

A third norm of price behavior is found in the equality between price and the socially relevant marginal cost of adding to industry output.[6] When such a condition holds for a number of industries at once, the relative outputs of the several industries are governed by a consistent rule and to some advantage. This matter will be discussed at length below.

It is also relevant that purely competitive prices in general are flexible or responsive to changes in cost and demand. For any one industry, price flexibility tends to make output more stable in the face of given fluctuations in purchasing power; for a number of industries it tends, when buying power fluctuates, to keep their relative outputs in a relationship governed by impersonal cost and demand conditions rather than by the relatively arbitrary decisions of individual enterprisers.

It is further significant that production takes place without any "selling cost" for advertising, sales promotion, and the like, since no seller has anything unique to advertise. Although the total absence of such costs is not necessarily ideal in every market situation, an absolute maximum of economy in this type of expenditure must be counted as a virtual advantage of pure competition. The significance of selling costs in other market situations must be discussed at a later point.

The progressiveness of firms in purely competitive industries, as measured by the readiness with which they adopt new techniques, improve products, or expand plant to supply the needs of a growing market, cannot be conclusively assessed on a priori grounds. On the one hand the freedom of entry to such markets, together with the impossibility of any firm withholding developments for fear of adversely affecting price, are definitely favorable to progress. On the other hand, the individual seller's complete lack of control over his market price may make any contemplated investment seem more risky than it would, for example, to a monopolist. Also, very small sellers may lack adequate funds for research. Both of these factors tend to dampen progressiveness. The net effect of pure competition on

[6] And, at the same time, the aggregate cost of producing the total output for which marginal cost equals price is the minimum attainable aggregate cost of that output (lowest available average cost per unit times the number of units), so that efficiency is at a practical maximum.

progress depends on the relative force of these two sets of factors. We will return to this matter in our discussion of monopoly.

The preceding appraisal of the performance of purely competitive markets stems directly from an exploration of the logic of profit-seeking within this sort of market structure. The conclusions are deduced rather than inductively discovered. The student is therefore entitled to ask how these conclusions square with observed results in actual markets.

We should first re-emphasize that very few markets in the United States have structures reasonably approximating pure competition. In mining, manufacture, transportation, and communications, concentration has with very few exceptions proceeded to the point where the number of sellers is small enough that individual sellers can influence the market price. Many of these markets are further complicated by differentiation among the products of rival sellers, which gives such sellers further direct control over their pricing. In the distributive and service trades, the number of sellers in a market is ordinarily large, but product differentiation is important enough to cause a distinct departure from purely competitive conditions. Agriculture has ordinarily been cited as the final stronghold of pure competition, since farmers are ordinarily many and their products highly standardized. But even here the growth of very large-scale farming in certain crops, coupled with the formation of producers' cooperatives in many others, has frequently resulted in the concentration of output in the hands of a few sellers. As a result we find anything like pure competition among sellers only in one of two industrial markets and in a part of the agricultural sphere.

In this limited area of the economy, purely competitive markets seem to behave about as our theory would lead us to believe. The average profits of enterprise are low—certainly lower than in many fields where strong monopoly elements are present. Prices tend to be very flexible over time, responding very strongly to cyclical changes in the level of buyers' demands, and the rate of output is correspondingly more stable over time than it is elsewhere in the economy. There is also incomplete evidence that output is generally adjusted to keep

marginal costs of production close to price. Scales and rates of utilization of plants evidence no consistent average departures from the optimum.

An extended perusal of an individual industry—such as the cotton gray goods or bituminous coal industries—which approximates conditions of price competition, may of course reveal special considerations affecting price which are not accounted for in a general theory of pure competition. Unstable wages or material prices, for example, have for a long time led to rapidly shifting costs and extreme instability of price in both the cases just mentioned. Price reductions once under way tended to stimulate further instability of wage and other costs—in effect competitive wage-cutting—with undesirable end results for labor employed in coal mines and cotton mills. Both industries, moreover, were plagued during the long interval between World Wars I and II with what appeared to be insufficient profits. Yet the process of exit from these industries, which should have tended to restrict output enough to raise prices to a normally profitable level, was very slow. The long run was very long indeed, if viewed as a period during which capacity might become adjusted to demand. Similar difficulties of accentuated price instability and of slowness in adjusting to declining demand have plagued agriculture and other purely competitive industries.[7]

Put in strict analytical terms, the peculiar difficulties of these industries include: (1) The fact that industry demand has declined over long periods, from wartime or other irregular peaks, leading to the need for progressive exit of a significant proportion of established firms if a long-run equilibrium were to be struck. (2) The fact that fixed plant—whether coal mines, cotton mills, or developed agricultural land—is long-lived, so that exit is very slow to take place so long as out-of-pocket costs are covered. For very long time intervals, therefore, firms continue to earn quasi rents less than full fixed costs, and their accounts show profits less than normal. (3) The labor supplies in these industries (including self-employed labor in farming) are

[7] See Lloyd G. Reynolds, "Cutthroat Competition," *American Economic Review*, December 1940.

immobile and have not tended to leave these industries as the demand for their services declined. As a result, the declining demand for the products has led to lower wages (and lower variable costs) rather than to smaller output. This has in turn delayed the exit of firms.

It is also true that sellers in purely competitive industries have been at an absolute or comparative disadvantage because so many industries in the economy were oligopolistic and highly concentrated. When this is the case, entry of potential new enterprisers into the concentrated fields (where large investments are ordinarily required) is likely to be quite difficult, and the few remaining industries with many sellers and easy entry are likely to become overcrowded, especially if there is any tendency to over-all unemployment. These industries therefore tend to realize smaller profits than those with more difficult entry, and sellers in them may feel that they suffer from "too much" competition. Not only is abnormally active entry a problem, but with some given number of sellers these many-seller industries suffer comparatively also because they cannot arrange an organized restriction of output when demand declines, whereas in concentrated industries such output restriction may be quite feasible. The depression-period earnings of unconcentrated industry are thus likely to be much less favorable than those of concentrated industries.

As a consequence sellers in such industries have generally been unhappy in their lot, have decried as destructive, cutthroat, and murderous the sort of competition to which they are subject, and have sought special relief from the rigors of competition. This drive to escape pure competition has been basically responsible for producer cooperatives in agriculture, for government price supports in the same field, and for special price regulation for bituminous coal. There is probably an inherent tendency in capitalist enterprise to attempt escape from pure competition by one device or another, and this in part explains why this sort of market structure becomes less and less important through time. This tendency is accelerated when a great number of industries have attained the greater stability and higher profits ordinarily associated with more concentrated market structures.

In spite of the several socially desirable properties theoretically associated with pure competition, therefore, there is some question that it has ever been a practically tenable sort of market structure for very many industries in the economy. This is partly because the free enterprise on which it depends tends in its own interest to alter this sort of market situation as much and as quickly as possible. It is also partly because pure competition in the markets for goods has fostered excessive price instability, which was reflected in underlying labor and raw material prices, and because this put the areas affected in a much more precarious position than the remainder of the economy.

Nevertheless, the functions of market price as a regulator in industries in pure competition gives us a good simple picture of the *general* function of price in any market. At the same time, it gives us a reference mark in terms of which pricing in other more common sorts of markets may be described and measured.

NORMATIVE PROPERTIES OF AN ECONOMY IN PURE COMPETITION

The function of a price system, however, will not be revealed simply by considering one industry at a time. Perhaps the most important task of the price system is to guide the allocation of resources among different industries in accordance with relative cost and with consumer demand. Analysis of this process requires that we examine a large group of industries (preferably all of them) at once. In the actual economy, of course, the various industrial markets are of divergent structural types. Some are monopolies; many are oligopolies; monopolistic competition is fairly important; and there are a few markets in pure competition. In this mixed situation the function of the price system is quite complicated. To understand certain basic elements of the price system, therefore, it is useful to extend our analysis of pure competition to consider briefly how an economy with all its industries in pure competition might function. This will give us certain simple general ideas which will be useful later on, and will at the same time pro-

vide some reference points for the measurement of more complex (and more real) situations.

Suppose that we had an economy in which every industry was purely competitive—where in each industry there were many small sellers with a uniform product, and where entry to all industry was free and easy. Suppose also that there was a given constant flow of money purchasing power seeking all goods, which was maintained invariant through time. The accomplishments of a free price system should be clear and definite.

We should first expect every industry in the long run to tend toward a purely competitive equilibrium—in each industry, price should equal both the minimum average cost of production and the long- and short-run marginal cost. This result should be reached in the following manner. For each product which is being produced or can be produced, there exists at any moment a given industry demand schedule, showing the schedule of money offers by all buyers for that good. For the family of all goods there exists an interrelated family of such demand schedules, reflecting the relative importance, in terms of money expenditure, which buyers attach to various goods. Enterprising sellers, in pursuit of a profit, will be attracted by these monetary demands to undertake in each line a productive effort roughly proportionate to the money demand, considerations of cost being taken into account, and to increase output in each line until some balance between price and cost is struck. In this way, the money demands of buyers and the profit-seeking activities of sellers interact to accomplish some organization of resource use relative to consumer needs. This is generally true of any free-enterprise economy, regardless of whether market structures are purely competitive or are in a considerable degree monopolistic.

With all markets purely competitive, however, the adjustment would be influenced by the facts that all firms were free to enter any field without impediment and that in no field could individual firms affect the working of market price. Entry of sellers into each industry, and shifts from industry to industry, could then proceed until the return to capital invested was equalized in all fields and until no industry offered more attractive profits than another. The pursuit of profits by enterprise

should thus be able to adjust the relative outputs in various fields to relative consumer demands *without impediment*, until the force of competition made the discrepancy between cost and price in each similar.

Any *general* discrepancy between cost and price throughout the economy should also be eliminated, by competitive bidding among enterprisers for land, labor, and capital, so that factor prices in general would be adjusted to commodity prices in general in such fashion that there were no excess profits left. The net result of all this is that the economy should at last tend toward a sort of general equilibrium of relative prices of such character that in every industry price would be equal to both average cost and marginal cost. It would also necessarily be true, with a money flow which was self-sustaining and factor prices which were freely adjustable, that all factors would be fully employed.[8]

As this situation was approached, an impersonal price system would have imposed a very definite type of regulation on all economic activity. As long-run equilibrium was attained in all industries, resources would be allocated among the production of different goods in such wise that price was everywhere equal to the marginal cost of firms and to the minimum average cost of production of all firms at the prevailing level of factor prices. Now in this situation the following may be argued: (1) The socially relevant marginal cost of increasing industry output is the increment to aggregate industry cost caused by an increment to industry output, *given the factor prices prevailing at the point of the increment.* This cost reflects the rise of aggregate costs due to additional use of real resources at the prevailing prices—*the money value of the real cost increment*—but not any added rise in industry costs due to any *induced* rise in the money prices of factors. (2) The industry average cost as defined represents this socially relevant marginal cost of industry output, or its marginal real cost. (3) When this marginal cost equals price in each industry, as it will in a purely competitive economy, the allocation of resources among the production of various goods is ideal from the standpoint of total consumer satisfaction. The

[8] See Chapter 10 for further discussion of this point.

argument proving these points and qualifying them, however, is perhaps too complex for an elementary treatment. We may content ourselves with a sketching of the general argument subject to a simplifying assumption.

Let us assume that factor prices do not respond to changes in any one industry output and that no factor is supplied in limited quantity or successively worsening quality as industry output increases. Then the average-cost supply curve of the competitive industry, the locus of the minima of all firm average-cost curves at successive outputs, is horizontal. The marginal cost curve showing additions to cost for additions to industry output is then in every sense the same as the average cost curve, and shows only the money values of the (constant) increments to real cost per increment of output.[9]

From this it follows that a dollar's increment to industry cost in any industry buys a dollar's worth of *additional* resources, and (given perfect factor markets) the same amount of resources in each case—let us say one hour of labor in each industry. Let us further assume that the last dollar spent by buyers on each good brings them equal satisfaction. Suppose marginal cost equals price in every industry, as it will in competitive long-run equilibrium. Then the resources acquired by virtue of the last dollar increment to costs in each industry, and therefore the last hour of labor in each case, yields an output representing the same (a dollar's worth) increment to buyer satisfaction.

When this is true, allocation is ideal in the sense that any shift of an hour of labor from one industry to another will reduce the aggregate of buyer satisfaction. If we were to shift a dollar's worth or an hour of labor from any one industry to another, decreasing the first output and increasing the second, it would add the same physical output to the second industry as the last previous hour of labor had (since costs are constant), but this output *would be worth less than a dollar to buyers* because of diminishing marginal satisfaction with increased supply. It would

[9] And, at the same time, the output of each industry for which marginal cost equals price is being produced at the lowest attainable aggregate cost of that output (which in pure competition represents the output times minimum average costs). Producing every output at the lowest attainable aggregate cost is essential if resources are to be ideally allocated and used.

subtract a unit of output from the first industry worth a dollar (or, with further subtractions, more than a dollar because of increasing marginal satisfaction with reduced supply). Any such shift would thus result in some loss of total satisfaction from all goods, and the original competitive allocation is ideal. This may also be shown where the industry supply curve is not horizontal, excluding certain special cases. A first property of purely competitive equilibrium pricing for the whole economy would therefore be an ideal in allocation of resources among uses. This attainment only becomes especially meaningful, however, if all resources are employed.[10]

Two other properties of a purely competitive economy are merely the extension into a larger area of results identified with any purely competitive industry. As all industries reached a competitive equilibrium, prices would everywhere equal average costs, and for the whole economy there would then be no excess profits to enterprise. Competition would squeeze out such an excess wherever it arose, with the result that the entire income of industry (except as appropriated by the government) would be shared by labor in wages, capital in interest returns, and land and resources in rents. (The manner in which the relative shares of these productive factors is determined must be discussed later.[11] Entrepreneurship would make no return, once equilibrium was reached, in excess of the normal return on its own labor and its investment.

A second result of this competitive equilibrium would be that in all industries, firms would be forced to attain optimum size and rate of use of facilities. For the whole economy this should mean that every output was reduced at the minimum attainable aggregate cost (fortuitously coincident with minimum attainable average costs). Efficiency is this at a practical maximum.

[10] Precisely if (1) there is no involuntary unemployment of factors which wish to work at the going rates of pay but cannot find employment, and (2) if no resources are discouraged from seeking employment by monopolistic pricing practices. The second condition should be observed in pure competition, but the first may not be, as we will see in Chapter 12. For a further analysis of competitive and ideal allocation, see Abba P. Lerner, *The Economics of Control* (New York, The Macmillan Company, 1944), Chaps. 1-9.

[11] See Chapter 10.

This tendency toward efficiency would be reinforced by the fact that no resources would be devoted to selling and sales promotion (although such an extreme condition seems highly imaginative).

Progressiveness in a purely competitive economy would certainly not be absent and might or might not be as great as under any other organization of capitalism.

Considering the advantages we have enumerated, it might seem that for a whole economy pure competition would represent a relatively ideal state of affairs. Subject to its strictures, the price system would function automatically to effect satisfactory allocation of resources, distribution of income, and efficiency of production. Why, then, don't we have a purely competitive economy, or insist on having it? The answer to this lies in part at least in aspects of the function of the whole economy which we have so far largely neglected.

In the first place, pure competition in actual markets can be a practical ideal only if it is a tenable alternative within the real economy. Any brief history of modern industrialism shows that pure competition is not generally feasible in present-day capitalism. One basic postulate of pure competition is the existence in every market of a very large number of small firms, *each of which has been forced by competition to attain optimum size or scale.* It is therefore implicitly necessary, for the maintenance of any reasonable approximation to pure competition in an industry, that the techniques of production be such that every firm can grow large enough to exploit every advantage of large scale or mass production, and can still be small enough, relative to the total demand for the product, that no firm controls more than two or three percent of the market. As soon as techniques develop in such a manner that the firms in an industry must become so large to attain all economies of mass production that the market can support only a few of them, pure competition is self-destroying and is replaced by a market structure with few sellers instead of many.

One aspect of the later phases of the industrial revolution was to make pure competition technologically untenable in most manufacturing and transportation industries. Great economies were promised to very-large-scale production, and this encour-

aged the growth of individual firms to the point where the number in any industry could no longer be large. By 1890, there were in this country few industries where pure competition would not simply have destroyed itself and have been replaced by some form of oligopoly. This is not to say that the great size attained by many firms, and in general the degree to which industrial concentration proceeded, was wholly or even principally explained by the pursuit of economies of large-scale production. But such a pursuit alone could certainly have carried us far enough on the road to concentration generally to wipe out atomistic (many-seller) market structures in most industry. Pure competition may thus be regarded as technologically untenable in most of the economy, because really large numbers of sellers are now inconsistent with efficiency.

A second condition essential to the existence of pure competition is probably equally inconsistent with the functioning of modern capitalism. This is the condition that in any industry the products of various sellers should be and remain identical and undifferentiated. To be sure there are some industries, including several in agriculture, where the opportunities for product differentiation are quite slight, and others where the pressure of informed buyers forces some sort of standardization. But in the production of most consumers' goods, there is a clear tendency for sellers to differentiate their products from those of their rivals and thus seek some advantage in amount or security of profits. They tend to do this just as surely as they tend to equate marginal cost to price or to seek more efficient scales of plant, and they have done it in the great majority of cases. As long as we are talking about free enterprise, therefore, it is probably illegitimate to assume the general possibility of product uniformity among rival sellers, either voluntary or enforced. Market structures involving product differentiation seem to be integral parts of the capitalist system in a very large proportion of markets. For this second reason also, therefore, it is not reasonable to contemplate purely competitive markets as practically attainable in many instances in the modern world.

We do not have a purely competitive economy, never have had it, and could not reasonably sustain it in the free-enterprise setting in which it is relevant. One might therefore inquire why

we have discussed it, and also why we have mentioned certain results associated with it as ideals of economic performance.

The justifications for discussing pure competition are quite clear. First, although it is oversimplified and otherwise removed from reality, the purely competitive economy gives us a simple model within which the general function of a free price system may be observed. Oversimplification and artificiality are thus virtues if they assist us in learning the elements of a complicated process. Second, a purely competitive world (or one not significantly different from it) seems to have been the implicit reference of most of those political economists who, from Adam Smith on, have argued for a *laissez-faire* (*i.e.*, hands off) governmental policy toward the economy. They justified nonregulation by referring to how a purely competitive economy regulated itself. It may be useful to recognize the nature of this mythical economy to which they referred, and also to see how it differs from the real economy of then or now.

Admitted that pure competition is not tenable for most industries in the modern world, however, it does not necessarily follow that we err in ascribing normative properties to results hypothetically associated with purely competitive industries or economies. We have rejected the idea that pure competition is ideal as a market structure, since it may be inconsistent with technology or with the fundamental character of enterprise. But it may still be true that certain *results* hypothetically associated with it can serve as norms or ideals for the function of any price system. If so, they will have to be further established in any realistic setting where they are imposed. For the moment the results ascribed to a purely competitive economy serve only as potential or tentative norms for examination in connection with real market structures.

Analysis of the purely competitive economy has had the further advantage of establishing, on the abstract level, a definite law of behavior for prices and outputs which an economy would obey under the guidance of free and unrestricted competition. This pattern will at least serve as a measuring rod for other sorts of behavior we may encounter. And it clearly poses an important question: If this is not the pattern of behavior im-

posed by the price system we have, what *is* that pattern of behavior, and to what ends does the enterprise system automatically govern itself?

We have now in part established the usefulness of a theory of pricing which points to certain properties for the long-run individual equilibrium of purely competitive industries or general equilibrium of purely competitive economies. It must be emphasized, however, that even if we admitted the assumption that the structure of all markets is purely competitive, the theory developed to this point would hardly embrace all significant processes of a free-enterprise economy. What we have done so far is to examine the manner in which, in the assumed situation, prices, outputs, and price-cost relationships would be determined if there were a given predetermined family of market demand curves for all products (reflecting a given aggregate level of purchasing power and given buyer tastes), and also given relationships of cost to output for all firms in every industry. Taking these determinants as given, we have reasoned our way quite legitimately to the conditions of long-run equilibrium for an industry or for the economy. In doing this, however, we have neglected certain phenomena which in practice hamper the functioning of any price system.

This neglect stems in part from our implicit assumption to this point that there is a given stable level of general purchasing power, from which the governing family of demand curves for various goods is derived. We must assume some such stable situation, and preferably at a level of full employment for the whole economy, if we are to suppose that any long-run price equilibrium is fully worked out. But in fact we know that the general level of income and purchasing power is continually on the move in a never-ending series of business fluctuations, and that any persistent stability at full employment is rare indeed. It is therefore clear that in practice we should not theoretically expect a purely competitive industry or economy to attain and hold long-run equilibrium. Rather we should expect it to proceed through an ever-changing series of short-run equilibria and to pursue some sort of long-run equilibrium—perhaps as some average of fluctuating short-run values. The results here-

tofore ascribed to long-run equilibrium should exist as inherent tendencies in an unstable and fluctuating situation.

It is also clear that we have not undertaken to explain the origin of the persistent fluctuations in general purchasing power to which we have referred. We need not necessarily do so at this point, so far as this is logically a separate task, concerned with the determinants of income, investment, consumption, and saving for the whole economy. So far as price behavior influences the stability of the economy and the character of fluctuations, however, this influence should be taken into account in appraising any sort of market pricing. Assuming then the preexistence of a tendency to persistent instability or fluctuation in the economy, would the pricing which characterizes a purely competitive economy in any way influence these fluctuations?

Extending our previous arguments from the single industry to the whole economy, it is clear that if all markets were purely competitive, the price system in general would be very responsive to fluctuations in income, with both absolute and relative prices reacting sensitively to changes in demand. It is not clear, however, that this would necessarily be an advantage to the economy. General price flexibility would *virtually* tend to stabilize output, provided that we could suppose that the fluctuation of purchasing power was itself uninfluenced by the flexible changes of price it induced. Then any given fluctuation of income, initiating shifts in the demands for various products, would effect a smaller fluctuation in output the more flexible prices were. But it is possible that any induced fluctuation in price in turn generates further fluctuation in income, and thus intensifies any initial fluctuation in purchasing power. It will do this if, with any initial price change, buyers *speculate* on further price changes by buying in advance if prices are rising or withholding purchases if prices are falling. In this event, price flexibility may be cumulative and may intensify instability. The case for a purely competitive economy in a world of dynamically fluctuating income is therefore by no means clear, and must be deferred to more detailed discussions of the matter. We will refer to the issue again in Chapters 9 and 10.

To this point we have examined the process of price determination and the regulatory function of price in industries in pure

competition and for a hypothetical economy in pure competition. We are now ready to turn to pricing in more complex situations.

SUPPLEMENTARY READINGS

EDWARD H. CHAMBERLIN, *The Theory of Monopolistic Competition*, Chap. II.

J. E. MEADE AND C. J. HITCH, *Economic Analysis and Public Policy*, New York: Oxford University Press, 1938, Part II, Chaps. 1-3.

ABBA P. LERNER, *The Economics of Control*, New York: The Macmillan Company, 1944, Chaps. 1-18.

GEORGE J. STIGLER, *The Theory of Price*, Chap. 9.

ARTHUR R. BURNS, *The Decline of Competition*, New York: McGraw-Hill Book Company, 1936, Chap. 1.

CHAPTER 5

PRICING IN MONOPOLIZED MARKETS

The preceding discussion of how an economy would work if all the markets for all commodities were in pure competition clearly poses some questions about the actual economy of the United States. Granted that our economy is not atomistic in structure, and granted that its behavior is probably not that ascribed to a purely competitive system of markets, how does it behave? Is there a law of behavior which the economy we have does obey? These queries can be satisfied if we can determine, for the sorts of markets which we have in fact: (1) how firms determine their prices and outputs; (2) how industry prices are made for groups of rival firms, and to what ends; and (3) how the resulting system of prices for all industry operates to accomplish its regulatory functions.

Any investigation of pricing in the real world is of course guided by the fact that a variety of market types is practically important, including monopoly, monopolistic competition, and several sorts of oligopoly. Each significant market category must be investigated separately. There is some advantage to beginning this investigation by considering simple or single-firm monopoly—the market in which one firm controls all the supply of a good and is not troubled by the competition of any very close substitutes. Such monopoly is, to be sure, not the most common thing in our economy, although there are important

instances of substantial monopolization of entire industries. But the sort of pricing which occurs in single-firm monopoly is in many respects a simplified and somewhat extreme version of pricing in the great mass of oligopolistic industries which dominate the economy. For in all these industries there is an element or "degree" of monopoly, which tends to be exploited in much the same way as a simple monopolist exploits his position. We can therefore learn most rapidly about several significant aspects of real pricing by considering first the determination of price and output by a single-firm monopolist.

Monopolistic pricing, like pricing in oligopoly or monopolistic competition, turns out to be different and more complicated than purely competitive pricing. The reasons for this are fairly evident. In the first place, the underlying market structures in monopoly and oligopoly are significantly different from that of pure competition. Instead of many competing sellers in a market, there are ordinarily a few, often a very few, and in the case of simple monopoly only one. Where there are several rival sellers in a market, moreover, their products are very often differentiated, and susceptible to further dynamic differentiation, or product variation. Both concentration of numbers and product differentiation tend to give the individual seller a degree of control over his own price, or *price jurisdiction*, which he does not possess in pure competition. His exercise of this jurisdiction may affect price behavior considerably.

A second difference between real pricing and the hypothetical variety attributed to pure competition is that in the real world the scope of closely interrelated "price" decisions which a firm must make is much greater. If we artificially suppose the firm to have a given and unchanging product, and one which is identical to that of many rival sellers, then we may reasonably suppose that his principal concern is the simple one of adjusting production costs (there being no selling cost) to the going market price for the product in a single (the present) time period. But in actual monopolistic and oligopolistic markets the decisions which the firm must make include (1) adjustment of current output in the light not only of production-cost variations but also of connected price variations, (2) variations of its product according to the effect of these variations on present and

future demand and on cost, (3) variations of selling costs according to their effect on demand, (4) adjustment of present price and cost in the light of their effects on future price and cost. In a word, the firm in practice has to devise, instead of a limited output decision, a complicated price-and-market policy embracing price, output, product, production cost, and selling cost in both present and future dimensions. The function of the resulting price system, and of the competitive mechanism from which it stems, is thus quite complicated. We can now begin to cut into those complexities by considering pricing in a simple monopoly.

THE INSTITUTIONAL SETTING OF MONOPOLY

Single-firm monopoly occurs when one seller sells a product for which there is no close competitor or rival. This idea is, of course, not as simple as words can make it seem. Is a monopolist still a monopolist if the product to which he has sole rights has close substitutes? Does a sole seller of Camel cigarettes have a monopoly if another seller has Chesterfields, or is the match monopoly in Ruritania really a monopoly if there are other sellers of mechanical cigarette lighters?

Logically, a seller has a distinct product to monopolize whenever his output is at least slightly differentiated from every substitute—so that at least some buyers consider it somewhat different from other products. Any firm with a protected design, package, or trademark is thus a monopoly in some degree. But it may still have competitors or rivals, in the sense that other products are such close substitutes that its price is very closely linked to theirs, or that changes in the prices of these substitutes will substantially influence its demand. When a monopolist has a product which is simply one of a differentiated group of many close but imperfect substitutes, he "has competitors" and is really a participant in "monopolistic competition." When the monopolist's product has a few important substitutes close enough that he and the sellers of these substitutes take account of each other's price reactions in setting their own—where the prices of their products are recognized to be significantly interrelated—the monopolist "has rivals" and is

essentially in an oligopoly situation. These are the common settings for "monopoly" today—monopoly plus competition or plus rivalry. But neither is "single-firm" monopoly. We reserve this category for the case where (1) there is no group of close substitutes for the monopolized product concurrent changes in the price of which will substantially affect the demand for the output of the monopolist and where (2) there is no one rival product the price of which will respond to the monopolist's price changes enough to affect perceptibly the demand for the monopolized product.[1] Single-firm monopoly is thus monopoly without effective competition or rivalry.

We should hasten to say that no firm has probably ever been *quite* a single-firm monopoly. But there are a number of instances where substitutes are distant enough and effective rivalry slight enough that conditions of single-firm monopoly are approximated. These would include aluminum-ingot production before World War II, nickel production, telephone communications, magnesium metal production before World War II, and others. Monopoly pricing is thus well worth examining for its own importance, as well as for the light it casts on quasi-monopolistic situations.

The essence of monopoly in general is the situation of a seller in a position where he can produce some product and can *prevent* any other seller from producing an identical substitute. The single-firm monopolist simply has a very strong situation of this sort—he is protected from the rivalry of sellers producing any reasonably close substitute for his product. Other sellers cannot enter his market or come close enough to it to set up a direct rivalry. From this it is quite evident that an essential aspect of monopoly is an effective *barrier to entry* in the way of potential rivals, for if there were not such a barrier there ordinarily would be rivals. This is what clearly distinguishes monopoly from pure competition, where there is necessarily little or no obstacle in the way of competitive entry. Some blockade of entry must almost certainly exist if there is a monopoly.

[1] Single-firm monopoly may also extend to include a seller with "rival" products the prices of which react to his only in a unique, limited, and definitely predictable fashion and thereby affect his demand in a similarly definite and limited fashion. See page 45, note 12.

This blockade is occasionally "economic" in character—that is, the conditions of production may simply be such that the entry of more than one seller is not attracted by profit possibilities. This may be so if the optimum scale of a single firm is so large as to oversupply the whole market, or if supply by more than one seller is inconvenient to consumers, as in the case of telephone service. In those instances we have what has been called "natural monopoly," and public price regulation has ordinarily been considered desirable.

The barrier to entry, however, is just as often institutional or legal in character. Three principal types of legal barriers to entry are patents, trademarks, and tariff import levies. The grant of a patent is explicitly a 17-year monopoly over the patented article, device, or process. A trademark is similarly a monopoly grant covering the package or description of a good. Tariff laws may exclude the entry of foreign outputs of certain goods, thus affording a degree of monopoly to the domestic seller or sellers.

A common institutional blockade to entry arises when a single firm manages to acquire all or most of a strategic raw material required to produce a product and thus is able to exclude competitors at will. This is essentially an exercise of the rights bestowed by the law of ownership or property. Where a monopoly occurs, therefore, we must recognize that it may very often be neither a chance phenomenon nor a natural occurrence, but the result of some aspect of the deliberately adopted legal and institutional framework within which business operates. Most monopolies, to be sure, are less than single-firm monopolies; they are tempered by the active rivalry of substitutes. But occasionally a blockade to entry can be secured which establishes a real single-firm monopoly.

PRICING BY A SINGLE-FIRM MONOPOLIST

The distinguishing character of pricing in single-firm monopoly is that the monopolist can arrive at his price policy without much concern about the prices of other outputs. He has a market all to himself to exploit more or less as he will. Let us investigate the general principles of such pricing.

As we have indicated, the monopolist must make a wider range of decisions than we have considered heretofore. But he must at any rate make one common to all markets—how much to produce of a given good and what to charge for it at any given time. We will first consider how this principal decision is likely to be made in single-firm monopoly.

For any seller, the simple price-output decision for a period can be conceived as resting on (1) the seller's demand curve for that period, and (2) his cost curves appropriate to that period. (His product and his selling cost we will for the moment take as given.) The single-firm monopolist's choice of a price and a rate of output, in either short or long run follows the same logic as that developed for a seller in pure competition. That is, the monopolist will try to select that rate of output which maximizes his profit, and this will presumably be the output for which his marginal cost is equal to his marginal receipts from sales.

The monopolist's position in this matter is unique only in the relationship which exists between the volume of his sales and the price he can receive for them. The seller in pure competition views his sales volume as having no effect on selling price. He can offer as much or as little as he will without influencing the market price for his good, because he controls only a minute proportion of the total industry supply. The monopolist, on the other hand, controls the entire supply of a good, and he will find that variations in his sales volume produce definite associated variations in price. In short, the more he produces, the lower the price per unit will be. In diagrammatic terms, the seller in pure competition faces a demand curve for his output which is simply a horizontal line at the going price, whereas the monopolist's demand curve is a negatively sloped industry demand curve, and shows a clear inverse relation of sales volume and price, as in Figure 24.

The single-firm monopolist, moreover, will view the position of this demand curve for his output as impervious to any price-output decisions he himself may make. Since there are no close substitutes for or direct rivals of his product, he can select any price on this demand curve without inducing any rivalrous reactions which would in turn affect its position. No "group ad-

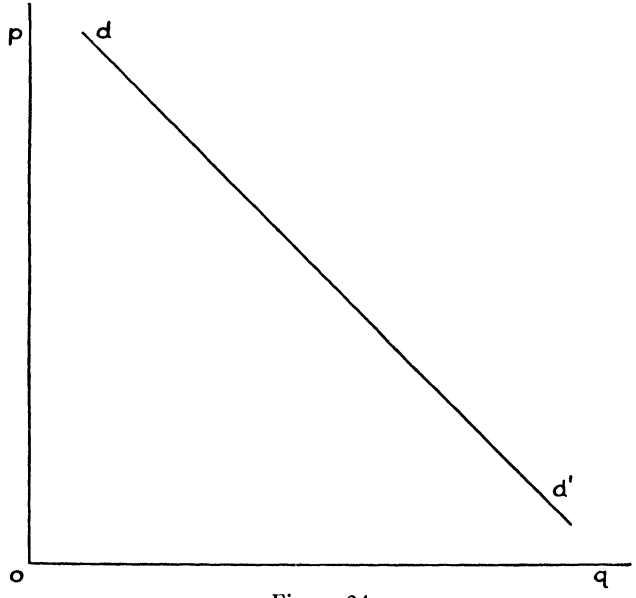

Figure 24

justment," other than the general equilibrium adjustment of the monopoly to the whole economy, need be taken into account.

Knowing or estimating the demand for his product, the monopolist can select the most profitable output by balancing the variation of selling price in response to variation in output against the corresponding variation in production cost. In effect he will set his marginal cost curve (let us say for the short run) against his demand curve, and will maximize his profits at the output where marginal cost is equal to marginal receipts—where the addition to total cost incurred in producing the last unit of output equals the addition to total receipts earned by selling it.

The marginal receipts from sales—the additions to total receipts for additions to output sold—follow a different course in monopoly than they do in pure competition. Since the purely competitive seller can extend his sales with no reduction in price (his demand curve is a horizontal line), any addition to output sold adds as much to his total receipts as the selling price of

that output. A sector of a seller's demand curve in pure competition might follow the pattern shown in columns (1) and (2) below. Then the marginal receipts should be equal to price, as shown in column (4).

(1) Price	(2) Quantity of output sold	(3) Total receipts (1) × (2)	(4) Marginal receipts (Addition to total receipts per addition to quantity)
$0.50	7	$3.50	—
0.50	8	4.00	.50
0.50	9	4.50	.50
0.50	10	5.00	.50
0.50	11	5.50	.50
0.50	12	6.00	.50

Marginal receipts are equal to price in pure competition, because the seller's price does not decline with increases in his output.

But the monopolist's demand curve slopes negatively—increases in his output sold are accompanied by reductions in his price. In monopoly the marginal receipts, or additions to total receipts for additions to output, are therefore always less than price. Thus if the monopolist's demand schedule is as shown in columns (1) and (2) below, marginal receipts are as shown in column (4), consistently less than selling price.

(1) Price	(2) Quantity of output sold	(3) Total receipts	(4) Marginal receipts
$0.50	7	$3.50	—
0.49	8	3.92	.42
0.48	9	4.32	.40
0.47	10	4.70	.38
0.46	11	5.06	.36
0.45	12	5.40	.34

When a monopolist extends his output—let us say by one unit—the price he receives for all units falls. The additional benefit or receipts the monopolist gets from one more unit are therefore not as much as its price, but less than that price by the receipts lost through falling price on the other units.[2]

In considering extension of his output, the monopolist is obviously concerned mainly with the marginal receipts from that extension, not with the price. The fact that in monopoly the marginal receipts of an additional output are always less than its price simply reflects a very essential aspect of monopoly pricing—because an increased output means a lowered price, there is always a virtual drag on or deterrent to extending output.

The marginal receipts resulting from any monopoly (or other) demand curve can be plotted as a curve derived from this demand. Where p is price and q quantity of output, the marginal receipts may be defined as $d(pq)/dq$; that is, as the rate of change of total receipts corresponding to change in output. Thus the demand curve dd' in Figure 25 has corresponding to it the marginal receipts curve mr. The curve mr shows the rate of increase of total revenue in response to increasing output sold. It will be clear that *the position of the marginal receipts curve depends on the elasticity of the demand curve from which it is derived*—the more elastic the demand, the closer to price will marginal receipts be. If demand is more elastic than unity, marginal receipts are positive; if demand is unit elastic, marginal receipts are zero; and if demand is less elastic than unity, marginal receipts are negative. The student should consider the preceding propositions until their logic becomes obvious to him.

Profit maximization for the monopolist (for a given product in a given period) stems from his choosing that output for which marginal cost of production equals marginal receipts from

[2] Thus in the preceding table, an eighth unit of output can be sold at a price of 49 cents. But it does not add 49 cents to the receipts of the firm, since in order to sell an eighth unit, the price of all units must decline from 50 to 49 cents. This entails a virtual loss of 1 cent per unit on the preceding seven units, or a loss of total receipts of 7 cents. Setting this loss against the added 49 cents for the eighth unit, we get marginal receipts of 42 cents.

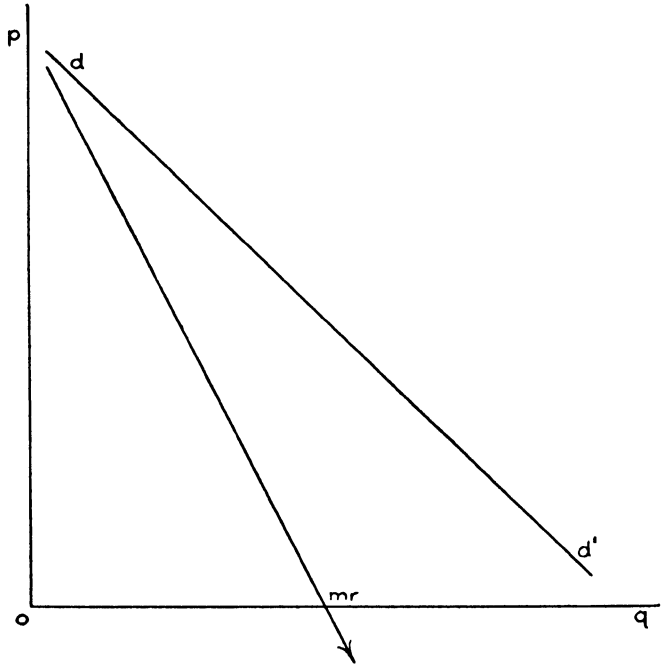

Figure 25

sales. His decision as to how much to produce and what to charge in a short period is charted in Figure 26, which brings together the monopolist's demand and marginal receipts curve and his short-run cost curves. Here *dd'* is his demand curve, *mr* the marginal receipts drawn therefrom, *ATC* the average total cost of production, and *MC* the marginal cost. (Average variable cost and average fixed cost curves are omitted for the sake of simplicity.) Assuming that the monopolist's motive is to maximize his profit for the short period, he will extend his output just so long as the added cost of additions to output is less than the added receipts they bring in sales. This will bring him into equilibrium at the output *oq*, charging the price *op*, where marginal cost equals marginal receipts. This output evidently makes the aggregate profit *abcp* the largest obtainable with this cost and demand. (The student should momentarily disregard the price p_1 and the quantity q_1.)

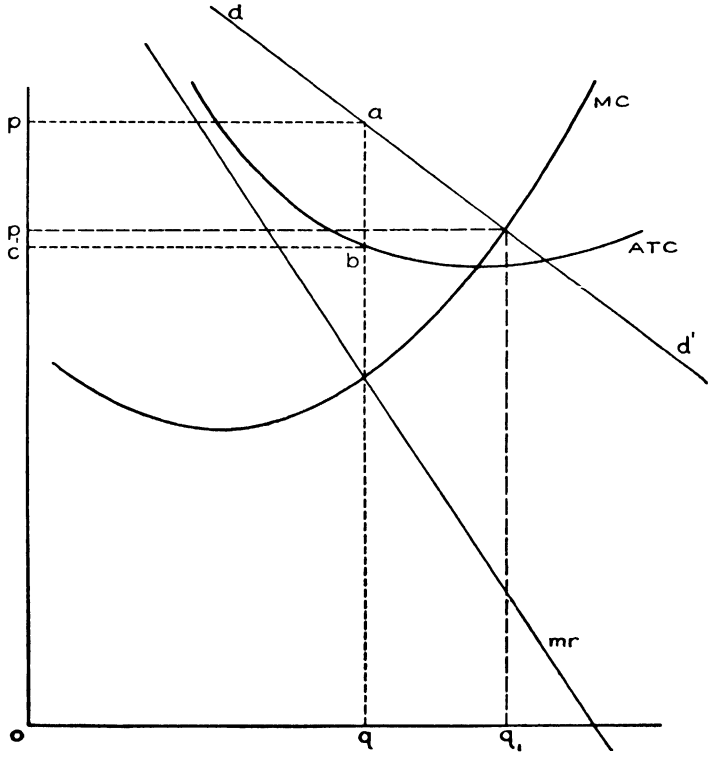

Figure 26

In this way the monopolist should decide, for a short period, (1) the price he will charge, (2) the output he will produce, and (3) the rate at which he will utilize his given fixed plant.

The monopolist's long-run adjustment of output to demand and price should follow strictly similar lines, provided we view his product as invariant and exclude selling costs. The monopolist should choose that long-run rate of output, and corresponding scale of plant, for which long-run marginal cost is equal to the marginal receipts expected over the longer period. If in Figure 26 we read all curves as applying to the long-run rather than the short-run calculations of the seller, the same solution applies without change.

Although our analysis so far is limited by certain arbitrary

assumptions, it should now be possible to detect some tendencies inherent in monopoly pricing. What sort of price results tend to emerge from monopoly, and how do they differ from those of pure competition?

MONOPOLY PRICE RESULTS AND THE GENERAL ECONOMIC WELFARE

The relationship in monopoly of price to the average cost of production, and the resultant size of profits, is one important matter. For the short-run, to be sure, the general *limits* on monopolistic price–average-cost relationships are the same as those for a competitive firm. Price, that is, may be above or below or equal to average total cost, so long as it exceeds average variable cost. The single-firm monopolist may find himself in a short-run position where net losses are inevitable and may still produce, even though he fails to recover all of his fixed or sunk costs. Thus the short-run relation of his demand to his cost does not have to be like that shown in Figure 26; it could be such that there was no output for which price was as great as average total cost. This does not mean, however, that a single-firm monopoly is just as likely to incur short-run losses, and no more likely to make short-run profits, than a competitive firm. There is certainly a greater disposition toward higher profits and prices in monopoly. If there is any market output at which market price exceeds cost and allows a profit, the monopolist is free to and presumably will choose this output. The short-run returns of a seller in pure competition, on the other hand, are at the mercy of a market price which always tends to be driven to the level of marginal cost. The monopolist can reject such a price (p_1 in Figure 26, the equivalent of short-run competitive price) [3] and can choose any higher price that allows more profit. Although the single-firm monopolist *may* sustain short-run losses, therefore, he is less likely to do so than any other sort of seller.

Long-run excess profits are also more likely to occur in single-firm monopoly. As in pure competition, the effective minimum

[3] That is, it is the price equal to marginal cost.

for long-run monopoly price is at a level equal to average cost. But there is no effective price maximum as long as the monopoly is maintained. In pure competition, with free entry, profits in excess of normal return to investment tend in the long run to be eliminated. In single-firm monopoly, with completely impeded entry, long-run excess profits can be as large as the relation of costs to demand allows. The monopolist is free to choose the price which maximizes his profit, and therefore will tend to arrive at a price greater than average cost if there is such a price. Monopoly thus has a clear predisposition toward excess profits.

It has sometimes been the fashion to refer to the excess profit of the monopolist (the area *abcp* in Figure 26) as a return to whatever it is he possesses which impedes competitors from entering his market and thus eliminating his profit. The excess profit might accordingly be called the earning power of his patent on a process, of his public franchise to monopolize a field, or of his trademark. Such a terminological venture is useful in calling attention to the institutional source of most monopoly earnings. When one proceeds along this line, however, to the extent of calling this return a cost (since it is the necessary return on the investment that would be made in the patent or franchise if the monopolist had had to pay all it was worth to him in excess profits), he is obliterating a valid distinction between costs and excess profits and simply confusing the issue. We cannot whistle away the idea of excess profits. Explaining the source of excessive earnings does not keep them from being excessive.[4]

There is, in sum, a tendency in monopoly toward profits in excess of the necessary rewards to invested capital and to management. Such profits are simply the earnings of artificial scarcity, imposed by a monopolist with the aid of some barrier to competitive entry which protects him. Against this must be set the fact that it is always possible to find an unfortunate monopolist who, even through the fullest exploitation of his

[4] At the same time it is necessary to distinguish between costs and excess profits by defining the monopolist's costs very carefully—generally as the market value of all factors, like labor and capital goods, employed, but excluding the market value of special barriers to the entry of other sellers. In practice this poses a complex problem.

monopoly power, is barely able to make ends meet. Monopoly may be a real necessity in an instance of this sort.

The argument so far is that in specific monopolized fields there may be a relation of demand to cost such that the monopolist can and will earn an excess profit. Fundamentally, this is equivalent to saying that the prices of productive factors in general—which the monopolist purchases and which make up his costs—are so related to the demand for the monopolist's output, and to the most profitable price he can charge for it, that he can earn an excess of commodity price over full cost. This result is quite likely if factor prices are determined largely by the bids of competitive industry where outputs are not restricted and commodity prices are lower. Then the monopolist, by restricting his output, can enjoy a high product price, set by himself, and a lower level of factor prices, set or dominated by many competitive industries. It is also a likely result if many or all industries are monopolized, and if at the same time there is a restricted entry of enterprise generally into *additional* fields or industries, so that there is not sufficient bidding for factors to bring factor prices and costs into equality with monopolistic prices.

The basic setting for monopolistic excess profits is thus a restriction in the supply of monopolistic enterprise bidding for scarce factors of production. This condition is found in fact in conjunction with monopoly generally. But it should be pointed out that an unlimited supply of enterprise—let us say in additional industries producing additional commodities—could force factor prices up sufficiently to eliminate monopolistic excess profits in all except especially favored industries. This brings us back to the practical significance of blockaded entry in connection with the monopoly phenomenon.

Part of the impact of monopolistic excess profits on economic welfare is fairly obvious. Because the recipients of excess profits, the shareholders in monopoly companies and perhaps their executives, are relatively few and wealthy, the addition of excess profits to their earnings tends to make over-all income distribution more unequal. The rich get richer and the poor get poorer. Another aspect of this same phenomenon is that consumers of the monopolized good pay for excess profits in the price of the

good—price is "high." And both the distortion of income distribution and the raising of price are the results of artificially imposed scarcity, made possible by some unproductive obstacle to the entry of competitors. On those grounds, monopolistic excess profits seem undesirable. The only possible avenue for justification of them lies in some function they may play in the progressive development of a free-enterprise economy.

Some writers have suggested that opportunities for enterprise to establish monopoly positions are useful incentives to the progressive adoption of new techniques and to the development of new products, and that whatever current distortions in income may be caused by monopoly are more than atoned for by the increases in output and efficiency which are engendered by the dynamic pursuit of monopoly gains. This may or may not be true. At any rate it is not logically necessary that the advantages of monopoly should outweigh the disadvantages.

A second aspect of single-firm monopoly pricing is that output always tends to be set at a level where marginal cost is less than price. That is, the additional cost of producing the last unit added to output is smaller than the price which buyers are willing to pay for that unit. (This is the natural result of setting output so that marginal cost equals marginal receipts—see Figure 26.) In pure competition, on the other hand, output is generally extended until the industry supply curve, which is the relevant industry marginal cost, is equal to price. From this it appears that if a competitive or a monopolistic industry facing the same market demand curve were to have the same marginal cost or supply curve, the monopolist would produce less.

This comparison is inadequate, however, until we have considered the definition of costs under the two situations. In pure competition, the long-run industry supply curve is not a firm's cost curve but the locus of the minima of the average cost curves of all firms in the industry at successive outputs. It shows the marginal cost of any increment to industry output given the factor prices prevailing at the point of such increment—the money value of the real resources added to produce one more unit of output. When we say that marginal cost equals price in purely competitive long-run equilibrium, therefore, we mean that the money value of the marginal *real* cost equals price. The

long-run industry curve determining supply in monopoly is the marginal cost curve of a firm—presumably a much bigger firm than one in pure competition. It reflects any rise in the firm's money costs due to unit increments in output. But if the prices of all factors are given to this industry-firm, and do not vary in response to its output variations, its marginal costs will, like the competitive supply curve, reflect only the money value of the real resources added to produce a unit of output—the money value of the marginal real cost. In this case, the marginal cost of a monopolist has the same significance as the relevant marginal cost of a competitive industry—both measure increments to real cost valued at going prices. In comparing monopoly and competitive output in this chapter, we may assume for simplicity that the monopolist's output variations do not induce responses in the factor prices he pays, so that his long-run marginal cost may be considered on a par with the long-run supply curve of a competitive industry as a measure of marginal real costs. In either case then, for example, a dollar of marginal cost would reflect an added real cost of, let us say, 1 hour of labor, assuming both industries paid the same wage rate.

It may be noted parenthetically that where the monopolist's factor prices increase in response to increases in his output, his marginal cost curve is not the same as the supply curve of a competitive industry which faces similar factor-price behavior, and an equality of monopoly marginal cost to price then does not indicate the same relation of real cost to price as when the competitive industry equates its relevant marginal cost to price. The effect of such an induced rise in factor prices paid by a monopoly will be discussed in Chapter 7. In this chapter we will assume that the monopolist pays constant factor prices regardless of output, and that his marginal cost reflects increments to real cost in the same way as the supply curve of a competitive industry.

In monopoly then, such a marginal cost is less than price, and in pure competition marginal cost equals price. Output would be extended if an industry were competitive until the money value of the increment to real cost for another unit of output was equal to price; if the industry were monopolized it would only be extended until marginal cost so defined was equal

to marginal receipts. Effectively, assuming that the industry in either case would pay the same factor prices and have the same real costs, the monopolist would employ fewer units of real resources and hence produce less output. Suppose that either a competitive or a monopolistic industry would have a horizontal long-run marginal cost curve showing a constant marginal cost of $1 or 1 labor hour per unit of output. The competitive industry extends output until price falls to $1, the level of marginal cost. The monopolist extends output only until price falls to, let us say, $1.25. The monopolist will thus employ fewer hours of labor and produce less output than a competitive industry would. Also, the price of the output added by the last hour of labor he employs will be higher than it would with purely competitive selling. This is the essential significance of the excess of price over marginal cost in monopoly.

The same comparison applies to the short-run outputs of competitive and monopolistic industries. If, in Figure 26, we regard the *MC* curve alternatively either as the short-run marginal cost of a single monopolist or as the summed short-run marginal costs (provisional short-run industry supply curve) of a large group of competitors, the comparison is clear. The monopolistic industry charges the price op and produces the output oq. The competitive industry with identical costs and capacity charges the lower price op_1 and produces the larger output oq_1.

It should further be noted that, whereas in those cases where a purely competitive structure can survive by giving unit costs as low as any other market structure could, the aggregate cost of producing any output for which marginal cost equals price is the lowest attainable, the same may not be true for an industry which is fully monopolized. When a single seller can produce more efficiently than two or more (there is a "natural" monopoly), then the monopolist's aggregate costs for his output are the lowest attainable, even though this output may not be such that average costs are at the "optimum" level. Then the restriction of output is fully reflected in the apparent discrepancy of price from marginal cost. But if the monopolist produces less efficiently than two or more sellers would, and excludes added entry by artificial barriers, then his aggregate cost for his output are not the lowest attainable. In this event, there are two dis-

crepancies from the sort of price-cost relationship attributable to pure competition: (1) there is an observed discrepancy between price and attained marginal cost, and (2) the aggregate cost of production for attained output is higher than the minimum attainable. Then monopolistic inefficiency is added to monopolistic output restriction.

The monopolistic restriction of output is clear. In practically all real instances, of course, the direct comparison of the relative merits of competition and monopoly in a given monopolized industry is idle because the whole level costs would be different if the industry were in pure competition. Ordinarily, the small-scale plants which would be consistent with an atomistic market structure would involve higher costs, and pure competition would not be economically feasible. We should not make the common mistake, however, of supposing that this fact vitiates our generalizations concerning monopolistic restriction of output. It may be true that a monopoly produces more output than a competitive industry would in the same line, since the cost economies of the monopoly may offset the monopolistic tendency to restriction. But it is always true that monopolistic output is *virtually* restricted to a smaller level than that attributable to a competitive industry with the same costs.

The relations of attained scale of plant and rate of use of plant to optimum scale and utilization pursue no such specific tendency in single-firm monopoly. With one firm monopolizing an industry, the size to which it builds plant and the rate at which on the average it uses plant will depend on the size of its market. Whether its scale and utilization are optimum, therefore, mainly depends upon how large the output of an optimum-size plant is in relation to the most profitable output. It is equally conceivable that the monopolistic firm might be too small, too large, or optimum from the standpoint of cost economies. The main point to be made is that, given an effective monopoly, there is no corrective tendency which inevitably drives the firm toward optimum scale and rate of utilization.

It should be noted, moreover, that it is not necessarily in the best interest of society that a monopolist should produce the output corresponding to his "optimum" scale. From a social standpoint, it is desirable first that any industry produce any

output it produces at the minimum attainable aggregate cost of that output, and second that it produce an output which is in desirable or ideal relation to other industry outputs.

On the first point, the monopolist may or may not attain the minimum attainable aggregate cost of the output he produces. He will if he has a "natural" monopoly, which means that one producer is more efficient than two or more would be. In this case, his departure from minimum average costs is socially undesirable only so far as it could be overcome in the course of adjusting total output to a more desirable level, as defined by the relation of marginal cost to price. He will not attain the minimum attainable aggregate costs of his output if he is less efficient than two or more sellers would be (having expanded well beyond optimum scale). In this case, at least a part of his departure from minimum average costs is eliminable at his attained or larger industry outputs, and so far as it could be reduced by having more and smaller firms and at the same time producing a desirable total industry output, it is socially undesirable.

On the second point, the monopolist will generally restrict output below a socially desirable level. If, in the course of its extension to a desirable level, average costs would fall, then this much of the discrepancy from minimum average costs is socially undesirable. But even if (1) industry output were moved to the ideal level, and (2) the number of firms were adjusted so as to permit lowest attainable aggregate cost of that output, the firm or firms then producing would not necessarily be able to produce at "optimum" scale, although it or they would possibly be closer to it than the single-firm monopolist acting in his own interest.

It follows that the "optimum"-scale average cost and output of a monopolist are not generally socially best (unless by sheer coincidence); some different output and cost ordinarily must serve as a criterion—in effect, socially ideal output, as defined in terms of ideal allocation,[5] and a unit cost consistent with the lowest attainable aggregate cost of producing that output. The monopolist ordinarily departs from this more precise standard,

[5] Cf. pages 164-165 below.

but not ordinarily in the same degree that he departs from optimum-scale cost and output.

The distinguishing character of single-firm monopoly pricing, as revealed by our simplified analysis so far, is that the monopolist has broad jurisdiction over his price and can thus have a price policy, setting price and output to suit his ends. This allows him to restrict his rate of output to his advantage, and thus to earn any excess profit that is at all available. The effect of such exercise of policy on output, price, and profit is clear and distinct.

The monopolist is enabled to restrict output and earn extraordinary profits by a blockade he can impose against entry by competitors—by a fence around his position which excludes close substitutes for his output. We have thus far been concerned with the static price-output adjustment through which a monopolist exploits a given protected position. Regarding the dynamic policy of monopoly through time, it is clear that enterprises will be continually concerned with maintaining the barriers against entry which protect them, and with creating and establishing new protected positions which will yield attractive profits. This is clear from the history of business, as well as from the logic of a profit-seeking system. Successive establishment and destruction of monopoly positions has characterized the progression of capitalist industries through time.

Returning to the static adjustment of the single-firm monopolist to a given situation, the question may arise whether ordinarily he actually exacts a price high enough to maximize his immediate profits. This issue is especially pertinent where the monopolist's demand curve is considered to be less elastic than unity over a wide range, so that it would seem that the smaller the output he produces, the more profit he makes. If a single seller had a monopoly in steel, for example, the demand curve for the product might appear as *DD'* in Figure 27, being less elastic than unity at least until a very high price was reached. In this case marginal receipts are negative in the observed range, and it would appear that the monopolist would choose the highest observed price, or one perhaps double or treble the cost of producing steel. Yet this conclusion is unreasonable, because the steel monopolist would undoubtedly hesitate to charge

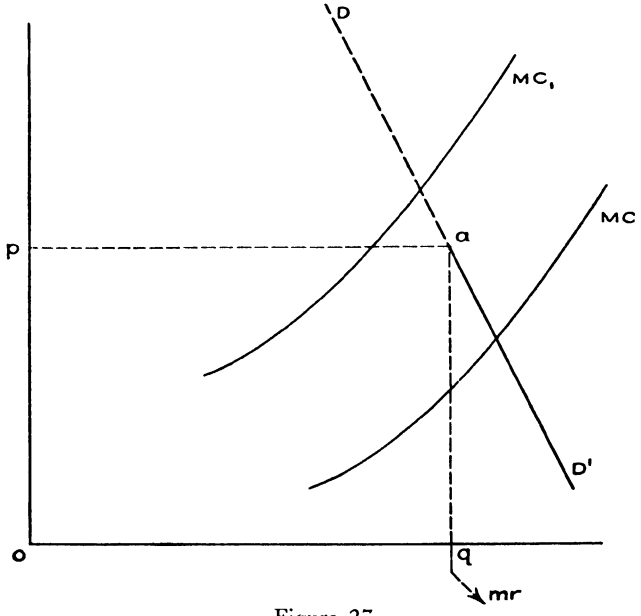

Figure 27

an obviously exorbitant price and to make a huge profit for fear either of attracting some rival firm to invade its field (in spite of all barriers) or of making buyers so indignant that public price control or other interference would be imposed.

In fact, therefore, the astute monopolist may consider some such price as *op* in Figure 27 to be his practical ceiling, above which he would attract competition or interference in spite of the barriers against entry he can place in the way of potential rivals. His own monopoly demand curve, for the long run, therefore, is really not *DD'*, but the *truncated* curve *aD'*. In effect, he can sell the amount *oq* at the price *op*, or larger amounts at the corresponding prices on *aD'*. But he cannot raise price above *op* without attracting entry or interference, and he cannot sell less than *oq at* the price *op* without having buyers bid the price—on resale—above this level. Marginal receipts are extremely uncertain above *op* and are negative beyond the output *oq*. In this situation the monopolist may logically produce *oq*, considering the price *op* the highest he can safely

charge, so long as the price *op* is equal to or greater than average cost at the output *oq*.[6] This will hold regardless of whether his marginal costs are below or above price (MC or MC_1). This practical limit on monopoly price is simply a reflection of the fact that no capitalist monopoly is absolute or impregnable, whatever legal or institutional protections it depends on. Any barrier against entry is only so high, and any monopolist's pricing power has a corresponding limit. In exploiting his monopoly, the monopolist has a problem similar to that faced by the government in taxing beverage spirits. If the per-gallon tax becomes sufficiently high, a substitute supply of bootleg spirits will appear in spite of the law and the revenuers.

ASPECTS OF THE DYNAMICS OF MONOPOLY PRICE POLICY

The preceding discussion has applied specifically to the adjustment a monopolist tends to make in the output of a given product for a single isolated time interval. Although this gives us some insight into monopoly pricing, it does not paint a lifelike picture of the operation of a monopoly through time. To produce a more realistic picture we must take account of (1) product and selling-cost adjustments through time, (2) adjustments of production technique, and (3) the interrelationship of pricing policies in a series of successive time periods.

Although we may have fairly characterized the tendency of single-firm monopoly pricing for a given product, it must be recognized that through time a product is not a static thing. It may be developed, varied, and changed in design. The monopolist, like any other producer, may find it possible to vary and develop his product to advantage. He may also find that variations in other products and the development of new products will continually affect the value of his own monopoly. An important part of his market policy will therefore involve employing commercial ingenuity and scientific research in developing his product, in order to perpetuate his monopoly and enhance

[6] Strictly, so long as the profit earned at this point is greater than he anticipates if he raises his price and attracts entry or interference. (Where profits are too low if entry is excluded, the monopolist may, of course, raise his price and take his chances on an industry with rivals.)

his own profit-making position. With any product he has at the moment, the monopolist will find he has a given demand curve, a given pattern of costs, and a given maximum potential profit. A *shift* in his product (through changed quality, design, or package) will have the effect of (1) changing the position of his demand curve, so far as buyers respond differently to the changed product, and (2) changing his costs of production, with the result that (3) he will have available a different maximum profit. The monopolist will therefore presumably follow a "product policy" which offers him, for some average of the future periods he contemplates, the largest profit. Thus in some cases he may be inclined to improve or raise the quality of his product, to "build his demand"; in others he may attempt to reduce price and cost to reach a great mass market.

Any seller who is able to do so, of course, is likely to adopt an active product policy, with similar profit motives. The single-firm monopolist, however, is in a singularly strong position to regulate the development of his product. He has no direct rivals, and therefore he is not likely to become involved in direct competitive product variation in order to maintain a competitive position. He can attempt to build his demand through product development without inducing competitive "product reactions" from rivals. Moreover, since he is the only producer of his sort of product, he can deliberately adopt or reject various potential product changes solely in terms of their profit prospect to him.

The institution of single-firm monopoly will probably mean that product development will follow a different course than it would if several sellers were to vie one against another for the same market. In effect, the monopolist will introduce a new product only if it is more profitable to him than the current product, whereas a new entrant to a free market will presumably introduce a new product if it is going to be profitable at all. Product change and development are thus likely to be slower in single-firm monopoly than in markets with free entry and effective rivalry.

The monopolist will also incorporate in his over-all price policy some expenditure on "selling," or promoting the sales of his product. Since he faces no direct rivals, he will not be-

come involved in direct competitive advertising and sales promotion expense in order to defend his market from immediate rivals. But he will have reason to undertake advertising simply to advance the position of his own demand curve through time—to build a larger market and therefore get larger profits. The presumed effect of a selling outlay is to shift the demand curve to the right, thus allowing larger sales at every price and possibly larger profits. The monopolist will presumably add to his selling outlays up to the point where the relationship of demand to all costs (production plus selling) is such as to permit a maximum profit.[7]

These selling outlays become a part of cost which must be covered by price; at the same time the output of the monopolized good may be expanded, possibly with economies in production cost. Ordinarily, however, advertising and other selling costs will be substantially smaller in single-firm monopoly than in industries supplied by rival firms, since competitive sales promotion cannot occur.

Production technique is also a dynamic variable through time, and one which the monopolist will control as a part of his general price and production policy, adopting or withholding potential changes in productive technique in accordance with their effect on costs and profits. As in the case of product changes, the single-firm monopolist is in a position to make deliberate choices without being forced by his rivals. Accordingly, he will be led to introduce cost-saving technical changes only when, output being given, the average total cost of production with new techniques is less than the average variable cost with existing plant (the fixed cost of which is sunk). In a purely competitive industry, on the other hand, new entrants to the industry will presumably introduce new techniques if only average total cost of production is less than the expected market price.[8]

[7] Technically, the monopolist's equilibrium is determined as follows: Let p be price, q quantity, c aggregate production cost, and s aggregate selling cost. Then $c = c\ (q)$, $p = f\ (q,s)$, and the monopolist solves for a maximum of $q \cdot p - c - s$.

[8] Not every seller in a competitive industry will move faster than a monopolist in introducing new techniques. But some will do so, and enough of them to lower price so as to give buyers the immediate advantage of the cost reduc-

Monopolies will thus be virtually *slower* in changing techniques than competitive industries, except so far as this tendency is offset by reduced uncertainty regarding monopolistic investment, or by more knowledge of techniques gained by large-scale research. But these offsets are not certain, and monopolies can be conservative as well as sanguine. It is worth noting, moreover, that the barriers to competition which surround a monopoly may protect undue conservatism or plain error from the losses it would suffer in a competitive market. In the net, single-firm monopoly once established probably acts as a drag on technical innovation.

When we consider not only the price-output adjustment which a monopolist must make, but also decisions as to product, technique, and selling cost, it becomes apparent that the monopolistic firm must have a complex and multidimensional price policy. Its profits are simultaneously affected by the price it charges, the product it produces, the amount it spends on sales promotion, and its production technique. Decisions affecting all of these matters are interdependent, so that in practice it adopts a composite policy with several parts. Characterization of the composite result which tends to emerge is difficult; results might logically differ from case to case. It is clear, however, that the introduction of further dimensions into monopoly price policy does not erase the tendency to relatively restricted output and high price. In addition our analysis indicates that selling costs in single-firm monopoly tend to be relatively low, and that the rate of product development and of technological change is probably slowed by fully centralized control.

In dealing with the single-firm monopolist's price-output decisions to this point, we have neglected one significant consideration affecting them, namely, the interrelationships among the succession of time intervals in which he makes short-run pricing decisions. As we have indicated, the monopolist in considering any one short time period (for example, a year) will

tion made possible by the new technique. The remaining competitive sellers will use out their old equipment (which now earns smaller quasi rents) until its variable cost dictates replacement. Nevertheless, new techniques should find their way into competitive industries sooner and affect the level of price more immediately.

choose the output and price at which his current marginal cost is equal to the current marginal receipts. If we extend this reasoning to his operations in a succession of years, we would expect him to follow this policy independently in each year as it occurs, adjusting the price and output of each period to maximize the profit of that period. This policy, however, would be consistent with the maximization of the aggregate of his profits over several years only if the profits of one period were unrelated to the prices and costs of preceding periods.

It is improbable that such a condition is observed in fact. Instead, it must generally be recognized (1) that the price charged for a good in earlier periods will affect the level of demand for the good in later periods, and therefore the profits of later periods; (2) that the costs of earlier periods, especially selling costs, may similarly affect the profits of later periods; and (3) that the monopolist will wish to maximize his aggregate or average profits over a succession of future periods and be willing to sacrifice immediate short-run profits for greater average profits over time.

Given these considerations, the price the monopolist charges at any one time may not fully exploit the current market, because price is set and cost incurred partly with an eye to their effect on the demand of future periods. Thus it has been noted in the case of the aluminum monopoly from 1910 to 1930 that the company often followed a relatively low-price policy (lower than would maximize current profits), and that this may have been done in order to make aluminum more widely known and thereby to cause the future demand for it to expand more rapidly than otherwise. Similarly, it has been suggested that, with the onset of business depressions, monopolists may not cut prices as much as current profit maximization would suggest, because they fear that an abnormally low price would "spoil the market" in later periods—cause future demands at high prices to fall off. Hence a monopoly may have relatively rigid prices over the business cycle.

In general we may suggest that the monopolist's price-output decision is in effect a series through time of related price-output decisions. It is only the combination of these decisions which might be expected to maximize the monetary earnings of the

firm. The single-firm monopolist, with his ability to control present and future prices, is in an especially good position to follow a calculated price policy which takes account of the relationship of present price and cost to future profit.

With so many variables affecting his profits—price, product, technique, and selling cost—the monopolist, being only human, may of course have difficulty manipulating all of them at once. Some observers have suggested that certain monopolists may make price decisions infrequently, adhering to a generally profitable price and giving principal emphasis to other variables. This is of course quite possible. It is also possible that if the estimated values of strategic variables are uncertain, the monopolist may forego detailed calculations and instead follow the arbitrary formulas or rules of thumb in setting price, output, product, or selling cost. When this happens, monopoly behavior may not be closely predictable by deductive methods.

THE IMPACT OF MONOPOLY ON GENERAL WELFARE—FURTHER REMARKS

It should now be convenient to look back and consider the presumptive effect of single-firm monopoly on general welfare in a free-enterprise economy. The total result of monopolization has several dimensions, including (1) its effect on price and output, (2) its effect on income distribution, (3) its effect on resource allocation, (4) its effect on the accumulation of selling costs, (5) its effect on progressiveness in technique and product development, and (6) its effect on stability. In each of these respects, what does the public get as the result of having a monopoly? This question may be asked first on the supposition that there are a few single-firm monopolies in the economy, the remaining industries being more or less competitive, and second on the assumption that all industries are monopolized—that there is a world of monopolies.

Let us first consider the case of a few single-firm monopolies in an otherwise competitive economy. Here a first result tends to be a relative restriction of outputs and raising of prices of the monopolized goods, so far as the monopolists have and exploit protected positions. This means in effect that the prices of the

monopolized goods are raised relative to those of competitive goods, that buyers therefore substitute the competitive goods for the monopoly goods to some degree, and that the outputs of the former are increased relative to those of the latter. Although there is a restriction of the output of monopoly goods, therefore, there is not necessarily a restriction of the aggregate output of the economy. If the resources excluded from employment in monopolized industries move freely to other industries, employment will not be impaired so long as the general level of money purchasing power and the general level of money factor prices coadjust to permit full employment. The primary impact of a few monopolies, restricting their own outputs, is thus really on the allocation of resources among the production of various products and on the distribution of income (through excess profits). It falls on aggregate output and employment only so far as immobility of resources among industries is encountered, so far as income distribution affects over-all employment, or so far as monopoly reduces the efficiency of production, as measured by the ratio of aggregate cost to total output, below best obtainable levels. Where there is significant immobility of resources away from a monopolized industry, there is also "monopsony" (buyer's monopoly), to be discussed in Chapter 7.

The conclusion that monopolized output is restricted must be tempered, however, in several ways. First, monopolistic price at any one time may be lower, and output larger, than that which would maximize immediate profits, perhaps because a monopolist is currently pursuing a low-price policy to promote future demand. In this event, price may at least for a time be lower than the competitive norm. On the other hand, price may be very high and output severely restricted in business depressions if because of long-run profit considerations the monopolist follows a rigid price policy. Neither of these manipulations of the behavior of price over time is at all possible in pure competition. Second, there may be monopolized industries where in spite of every exercise of monopoly power, full costs can barely be recovered. In this case a competitive industry might not be able to survive at all, and monopoly pricing may be necessary to secure any output. Where this is true, the charge of output restriction

and high price is not valid, since these are essential to securing any supply.[9] Third, the monopolist may exploit his position only in a mild degree. That is, he may view the economic "ceiling price," above which he will eventually attract public interference or private competition, as relatively low; then the degree of output restriction and price raising is likely to be quite moderate. None of these qualifications, however, entirely obliterates the basic tendency of single-firm monopoly to promote a significant degree of output restriction and price raising.

By restricting outputs in the monopolized industries, monopoly not only shifts but distorts resource allocations for the economy as a whole. Where some industries are subject to those restrictions and others are not, or where the degree of monopoly restriction differs among industries, this results in a distortion of resource allocation among industries. In some industries (the "less monopolistic" or the competitive) production is extended until marginal cost is at or near price, whereas in the more monopolistic industries production is restricted to a point where marginal cost is farther below price.[10] Suppose that a given increment to the money cost of any industry may be taken as representing in each case the same increment to real cost in resources used, as it may be if all industries pay the same prices for resources and if marginal cost measures the money value of marginal real cost. Then the last increments to real cost expended in the more monopolistic industries are producing goods worth more to buyers than the outputs realized from similar final increments in less monopolistic industries. In this event, buyers would get goods yielding them greater total satisfaction if more were produced of the goods of the more monopolistic industries, and less of the others.

This may be illustrated as follows. Suppose there is a purely competitive industry producing good A, where the marginal cost (of the last unit produced) is $1.00. The price is also necessarily

[9] Unless output is to be subsidized by the state. This raises issues into which we will not enter here.

[10] The discrepancy tends to depend upon the elasticity of the single seller's demand. The less elastic his demand, the greater the discrepancy between marginal cost (= marginal receipts) and price. We must also allow, however, for potential monopolistic inefficiency referred to on pp. 152-153.

$1.00 for good A. Suppose also there is a monopolized industry producing good B, and the marginal cost (of the last unit produced) is also $1.00. But since in monopoly equilibrium price exceeds marginal cost, the price of B will be higher—say $1.25. Now the last dollar of cost (or the amount of labor and other resources which $1.00 will buy) expended in industry A produces goods worth $1.00 to buyers, whereas the last dollar of cost (or equivalent resources) in industry B produces goods worth $1.25 to buyers.

If all resources are already fully employed, it would evidently be to the advantage of buyers if resources were shifted from industry A to industry B, since a dollar's worth of resources *shifted* to producing one more unit of B will increase satisfaction by about 1¼ dollars' worth, whereas it will reduce satisfaction, through the loss of one unit of A, by only about a dollar's worth. *The shift could continue to the advantage of buyers until the ratio of price to marginal cost was the same in industry A as in industry B.* This point would be reached after a determined amount of shifting resources from A to B, since price would fall with increasing output of B and price would rise with declining output of A.[11] But the shift will not take place under the unregulated pursuit of individual profit.

It thus appears that when monopoly exists in some but not all industries, or when different degrees of monopoly exist in different industries, resource allocation is not such as to give maximum satisfaction to buyers.[12]

[11] To obtain even relatively ideal results, of course, monopolistic inefficiency which makes the aggregate cost of any relevant output higher than attainable should also be eliminated. See pp. 152-153.

[12] It is clear that when marginal cost equals price in some industries and is less than price in others, allocation is distorted from the ideal. And it is also clear that when marginal cost equals price in all industries, allocation is demonstrably ideal. (See p. 129.) But it would seem that if *alternatively* price exceeded marginal cost in all industries by the same proportion, allocation would also be ideal; if all industries were subject to equal degrees of monopoly (equal proportional discrepancies of price from marginal cost), allocation would be equally desirable. This is true provided the level of employment which resources seek to attain is not influenced by all-around monopoly. If monopoly forces resources into idleness by reducing their real wages and other rewards, however, then the allocation of resources as between employment and idleness

The effect of monopoly on income distribution has already been made clear. There is a definite tendency toward the earning of excess profits. As these are earned, more income goes to the relatively few recipients of business profits, and income distribution becomes less equal. The widespread occurrence of effective monopolies would seriously distort the distribution of income in an economy. It is also evident that monopoly may cause unwarranted departures from desirable efficiency by restricting output in such fashion as to raise unit costs and by producing given outputs of more than their minimum attainable costs. The extent of this departure from ideal, however, is not uniquely predictable.

The effect of monopolistic organization of industry on the accumulation of selling costs cannot be predicted precisely. With single-firm monopoly, however, we may ordinarily expect to get some selling expenses, either aimed at promoting general growth of the product's demand, or of an "institutional" character, possibly to build political good will for the monopoly. Monopoly will thus devote some resources to selling, whereas pure competition uses none for this purpose. But single-firm monopoly will tend to incur smaller selling costs than rival firms in industries with differentiated products where active selling and advertising competition develops. Serious wastes through competitive advertising will ordinarily not be chargeable to single-firm monopoly, as they may be to certain cases of oligopoly and monopolistic competition.

To this point, the appraisal of single-firm monopoly has been unfavorable. If at any one time we view the performance of a monopoly, it appears to be responsible for restricted output, high price, unnecessary distortion in income distribution, and unsatisfactory allocation of resources among uses. In each of these respects, it falls short of an attainable norm or ideal of behavior. The relation of monopoly to the degree of *progressiveness* in an economy offers a similar though less conclusive case. On the

is distorted, even though the allocation of employed resources among various lines of production may be ideal. In short, equal degrees of monopoly in all industries may permit ideal allocation of employed resources, but not an ideal offering of resources for employment. Price–marginal-cost equality throughout gives preferable results. Cf. Lerner, *op. cit.*, Chap. 9.

one hand, a single-firm monopolist will presumably tend to introduce technical changes and new products more slowly than they would be introduced in an industry with free competition and easy entry, provided that the knowledge of new techniques and products and the firms' allowance for riskiness of new investment is the same in the two cases. On the other, a monopolist may have a more secure market than a competitor, and thus feel that investments are less risky; also, he may have larger profits to spend on research, and thus become acquainted with new investment opportunities more rapidly than would small competitors. On balance, we cannot say a priori whether monopolies are less or more progressive than competitive firms, although it is clear that monopoly will protect a nonprogressive policy.

The matter, however, cannot be left here. It may well be possible, within the broader logic of capitalism, that the lure of monopoly profits is generally necessary to induce enterprisers to make numerous innovations. Any innovation involves a risk to the innovator—he cannot be sure, when he begins, that he will be able to recover the investment in the new technique or new product. It may well be that enterprisers in general are willing to undertake such risks because they believe that if they are successful they will be able to enjoy excess profits over an extended period—that they will be able to secure monopolies (complete or partial) in their new positions. This is one of the justifications for the monopolies granted under the patent law, and it is not wholly without point. Some writers have argued that monopolization in some degree is an inseparable part of progress in capitalism, and at least in the past was justified for this reason. For our present analysis, we must recognize that although a firm which already has a monopoly may not be especially progressive, the chance *of becoming a monopoly* may be a necessary lure to progress by profit-seeking enterprises.[13]

In the last two paragraphs we have drifted progressively toward consideration of monopoly not as an isolated but as a general phenomenon. Let us now explicitly inquire into the pre-

[13] See Joseph A. Schumpeter, *Capitalism, Socialism, and Democracy* (New York: Harper & Brothers, 1942), Chaps. 6-8, for a development of the viewpoint referred to above.

sumptive effect on (1) income distribution, (2) total output and employment and (3) economic stability, of the organization of an economy entirely by monopolies—of having all industries controlled by single-firm monopolies. This supposition is extreme, but it may offer a simplified approximation to the real economy.

Income distribution in an economy where most or all industries were monopolized—or behaved as if they were—would tend to include a substantial excess-profit share for enterprise in general. This is because every industry would tend to restrict its output and its bids for productive factors in such a fashion as to open up such a profit gap if possible, and because this would tend to depress the level of wages, rents, and interest relative to commodity prices in such fashion as to allow all-around excess profits. In this connection it must be remembered, however, that all the various monopolized industries would be indirectly in competition for productive factors, and would tend to bid their prices to a certain level. If now there were freedom of entry for additional enterprise in either occupied or new fields in the economy, the pursuit of excess profits by this additional supply of enterprise would increase the demand for factors and bring their prices up relative to commodity prices in such wise that excess profits would tend to vanish.

The tendency toward recurrent excess profits in an economy of monopolies thus rests basically on widespread barriers against entry which restrict the opportunity for new enterprises to set up in competition with a limited number of established enterprises. It should be noted that the preceding argument supposes that productive factors are bought in competitive markets by many buyers (many different monopoly industries) and sold by many sellers. Where there is monopolistic selling of labor or other factors, or concentrated buying thereof, the problem is more complicated and will require special treatment.[14]

The effect of all-round industrial monopoly on the aggregate level of output and employment for the economy is closely linked to the income-distribution issue. At the outset it should be emphasized that the fact that one monopoly tends to restrict

[14] See Chapter 7.

its output relative to that of more competitive industries does not prove that an economy of many monopolies therefore restricts aggregate output and employment. It is, to be sure, evident that if (1) the total flow of money-purchasing power, which determines the general level of money demands for goods, and (2) the level of all money factor prices, which determines the positions of all cost curves, are both given and fixed, *then* the imposition of monopolies would tend to result in smaller output and employment, higher commodity prices, and larger profits than if all industries were competitive. In this case (with given purchasing power and factor prices) if competitive pricing would just result in full employment, for example, monopolistic pricing would result in less than full employment. The reason that this conclusion cannot be made general is that we cannot carry over to the whole economy the assumptions which are appropriate to a single industry—namely, that total money demand is given and that money costs are given. For the economy as a whole it is quite possible that an aggregative adjustment of money factor prices to total money demand will take place that will eliminate any unemployment which general monopolization tends to create, although at the expense of excess profits or other compensating effect.

This may be made clear by a simplified example. Suppose that for the economy as a whole there is a given constant flow of money purchasing power, or aggregate demand, for all goods and suppose that each money factor price is also given and constant. Suppose that every industry is a monopoly and the number of such industries (and firms) is also fixed. Now in this situation the total output is supposed to be such that not all factors are employed—the addition of the individual industry outputs, each determined by balancing a given cost curve against demand, is less than maximum output and requires less than all factors. But now if all money factor prices will fall—reducing costs—*while at the same time the flow of money purchasing power remains constant*—all the monopolies will produce more, and a certain fall in money factor prices will produce full employment. The process would simply involve a sufficient shift in income distribution from other shares toward profits to employ everyone—the impact of monopoly would fall upon reduced

wages, rents, and interest rather than upon reduced employment. (Employment would fall only so far as fewer factors *chose* to work for reduced real incomes.) This could in fact happen if (1) money factor prices would adjust downward freely, and (2) total purchasing power would be sustained by virtue of profit receivers spending their increased incomes as rapidly as the wage earners from whom they were taken away. There is, then, an open possibility that the tendency to output restriction by each monopoly will mean not over-all output restriction but rather an over-all reduction in the share of total output received as income by the hired factors of production.

The provisos subject to which this is true, however, suggest the conditions under which monopolization will restrict total output and employment. If the money prices of hired factors are inflexible and will not adjust downward under pressure of unemployment, while money income will not expand, then monopolization can create or increase unemployment. Also, if a decline in money factor prices under pressure of unemployment leads to a decline in total purchasing power because the resulting addition to profits is not spent but hoarded, then also monopoly will tend to create or increase unemployment. The solution thus depends upon complex considerations governing the spending of income, and will be treated in Chapter 12. On the same supposition under which a competitive economy always yields full employment—that is, a self-sustaining constant flow of money purchasing power and freely adjustable money wages, interest, and rents—a monopolized economy also yields full employment, or the opportunity to work for everyone who wishes to. The difference is that the monopolized-economy equilibrium would involve lower prices for hired factors and higher profits, and also the loss of some factors from the quantity available for employment because of the lower rates of pay.

The relationship of monopoly to economic stability is not at all simple and clear. Most economic analysis would lead us to believe that economic fluctuations occur more or less regardless of the sort of market organization which an economy has. That is, we would have business cycles, with intermittent unemployment and depression, either in a world of purely competitive markets or in a world of monopolies. Instability is not peculiar

to certain sorts of market organization but is characteristic generally of capitalism. The main issue is whether a certain form of market organization, such as monopoly, tends to affect the course of economic fluctuations in some systematic way.

It is difficult to point to any systematic relation of monopoly to business fluctuations, except for two tentative conclusions. First, monopoly prices tend to be more rigid than competitive prices over the course of business fluctuations. Monopolists can control their prices at will in the face of shifting demands for their products, and ordinarily find it to their long-run advantage to change prices infrequently and by relatively small amounts. As a result of price rigidity over the cycle, fluctuations of monopoly output may be accentuated, and the impact of shifting money income on the real output of the economy may be more severe. On the other hand, price rigidity may minimize the *cumulative* effects of rapidly shifting prices, which occur when individuals speculate on changing prices. The net effect of monopoly on the severity of fluctuations is thus difficult to appraise.

A second effect of monopoly on stability stems from its impact on income distribution. So far as monopolies create excess profits, they make income distribution less equal, and divert more of the total income of the economy into the hands of relatively few wealthy individuals. The spending of such large incomes is only in part for ordinary consumption; a large portion is spent on investments. Investment spending is much less stable over time than consumption spending, and much more susceptible to fluctuation. It may thus be argued that high-profit monopolies indirectly dispose the economy toward spending habits more susceptible to disruption by periodic boom and decline than would obtain if income were more equally distributed. It is not suggested, however, that fluctuations would be eliminated in an economy with no monopoly profits.

So much for the general significance of monopoly for the welfare of the economy, at least so far as it can be determined by exploring the logic of monopolist's price and output policies. The conclusions we have arrived at concerning how monopolies act, of course, have an abstract basis. They are arrived at by describing the relevant market conditions faced by the monopo-

list in a series of assumptions, and then by deducing the course under these conditions of monopoly action directed at maximizing profits. It is interesting to inquire whether these conclusions are supported by evidence of actual behavior by monopolists.

EMPIRICAL EVIDENCE OF MONOPOLY BEHAVIOR

One of the more complete studies of a single-firm monopoly concerns the aluminum industry, where from before 1900 and until 1941 the Aluminum Company of America (Alcoa) had a substantial single-firm monopoly in the production of virgin aluminum ingots for the United States market.[15] This monopoly was at first secured by a patent on the basic extraction process, which expired in 1909. Thereafter, monopoly was maintained largely by Alcoa's substantial control of the necessary ore reserves, and by an alert policy in dealing with potential new entrants to the industry. Foreign competition was excluded up to a certain margin by tariff, and otherwise (it has been alleged) by international agreements among aluminum producers. In any event, the monopoly continued secure until the outbreak of the present war, when the government undertook a major expansion of aluminum-producing facilities to meet military needs.

The price policy followed by Alcoa over a long period of years seems to have been consistent with what theory would lead us to expect from a monopoly. Aluminum ingot prices were on the average high enough to yield excessive profits over a long period. On the other hand, they were not always high enough to exploit Alcoa's current position fully; during the 1920's, Alcoa apparently lowered price and sacrificed immediate profits in order to promote the growth of demand for aluminum, and thus to enhance future profits.

The scale and rate of utilization of Alcoa's plants seemed in general to be consistent with maximum efficiency. The company was by 1930 considerably larger than the minimum size necessary to realize all economies of large scale. Whether it had become large enough to encounter serious diseconomies of large-

[15] See Donald H. Wallace, *Market Control in the Aluminum Industry*, Cambridge, Mass.: Harvard University Press, 1937.

scale management was not evident. Aluminum output was virtually restricted so far as price was higher than marginal cost.

Alcoa's monopoly was, of course, not complete. As any monopoly must, it had its limits. For aluminum ingot, it faced the growing competition of aluminum scrap metal, which is suitable for some but not all of the uses to which aluminum is put. Some of the products made from aluminum, moreover, which Alcoa produced in integrated operations, faced effective competition by substitutes, whereas others did not. Aluminum electric cable competed with copper cable, but aluminum alloys for aircraft had no effective substitutes. This led Alcoa to pursue a pricing policy common to monopoly and also theoretically predictable—namely, discrimination among the prices charged to different classes of buyers. The aluminum contained in electric cable was sold at a lower price, competitive with copper, whereas the aluminum in some other products was sold at a higher price.

In theory a monopolist would ascertain the separate demand curve for his product by each of two (or more) classes of buyers—*e.g.*, the demand for aluminum in cable and the demand for aluminum in aircrafts alloys. A separate price would then be set in each market, depending on the position and elasticity of its demand curve. Markets with less elastic demand curves would be charged higher prices, and those with more elastic demands would be charged lower prices. Output would be divided between the two markets so as to maximize total profits.[16] Alcoa seems at least to have moved in this direction in discriminating among various classes of aluminum buyers.

The tendency of monopolies to strive to protect their positions from the intrusion of substitutes is seen in Alcoa's policy toward magnesium, an effective substitute for aluminum in many uses. Together with a foreign company, Alcoa is said to have long had indirect control of patents covering the processes whereby magnesium is extracted, alloyed, and made into a structural metal. These patents were licensed exclusively to a single company for its sole use. The price policy on magnesium was such that the magnesium price was sufficiently above the aluminum price that magnesium offered no serious competition

[16] For a discussion of monopolistic discrimination of the sort described, see Boulding, *op. cit.*, pp. 540-549.

to aluminum in its major uses. Magnesium production remained insignificant. This situation was altered when the government undertook to stimulate magnesium production during the war emergency.

SINGLE-FIRM MONOPOLY IN THE PUBLIC UTILITY FIELD

In the aluminum industry we find a fairly good example of the behavior of an unregulated single-firm monopoly. Most such monopoly industries in our economy, however, are not free to pursue unregulated price policies but are subjected to public price regulation. These are principally firms in the electric utility, gas utility, water supply, and local transportation fields, which enjoy local monopolies in most areas, and in telephone and telegraph communications. In these industries it has generally been accepted that competitive supply results in a less satisfactory service to consumers, or that competition is self-destructive and unstable to no good end. Since the supply of basic consumer necessities is also involved, public authorities (usually state or local governments) have declared such industries to be "natural monopolies" and have helped nature along by granting monopoly franchises to firms in the various localities. In turn, the rates of such utilities have been subjected to regulation by public commissions in order to insure consumers of reasonable prices for the utility services. Type and quality of service are also regulated. Rate regulation generally includes establishment of a "fair" general level of rates, and also fixing the pattern of price discrimination among various classes of users.

Monopoly price behavior under regulation may thus be substantially different from that ascribed to unregulated monopoly. Precisely the sort of behavior which is obtained, however, depends strongly upon the attitudes and abilities of regulatory bodies, and upon the restrictions placed upon regulation by the courts in protecting the regulated monopolies from "confiscation of property without due process of law." Characterization of the behavior of price and output in regulated monopolies would require much special study. It seems fair to say, however, that not all of the pricing tendencies inherent in monopoly

are overcome by regulation as it exists, and that purely competitive pricing certainly does not result.

The study of price regulation, however, lies outside the field of this volume. We therefore turn next to the problem of the behavior of groups of monopolists.

SUPPLEMENTARY READINGS

KENNETH E. BOULDING, *Economic Analysis*, Chaps. 24-26.

R. F. KAHN, "Some Notes on Ideal Output," *Economic Journal*, vol. 45, pp. 1-35.

JOE S. BAIN, "The Normative Problem in Industrial Regulation," *American Economic Review*, Supplement, March 1943.

CLAIR WILCOX, *Competition and Monopoly in American Industry*, Temporary National Economic Committee, Monograph No. 21, Washington, 1941.

JOSEPH A. SCHUMPETER, *Capitalism, Socialism, and Democracy*, New York: Harper & Brothers, 1942, Part II.

CHAPTER 6

PRICING AND PRICE POLICY IN OLIGOPOLISTIC MARKETS

The single-firm monopoly is important in the public utility field and has had some significance in mining and manufacture. It does not, however, constitute the typical market structure for the modern American economy. In most of our industrial markets, sellers possess a degree of monopoly, but they are not single-firm monopolists in the sense that their products have no close substitutes and that they have no recognized rivalry with other sellers.

The typical industry structure in our economy involves a "group of interdependent monopolists." Where such a group exists, each of a number of sellers ordinarily has a "degree of monopoly" in a product which has distinct characteristics and which the seller can protect from exact or close imitation. At the same time, the products of the various sellers in the same group are rather close substitutes to most buyers, like various makes of automobiles. The demands for the products of various individual sellers are thus rather closely interdependent. A third characteristic which is common is that the established group of sellers often *shares* a "monopoly position," in the sense that there are barriers protecting all of them from the entry of outsiders to the group. A group of monopolists may exist with less than all of those conditions present, but often all three occur together.

This "group of monopolists" market category is very large, and necessarily breaks down into several subclasses. At an earlier point we have already mentioned three:

1. An industry or group of sellers with distinct but close-substitute products, who explicitly recognize the interdependence of the prices of their several products, and thus have a recognized rivalry. (A price change by any seller can elicit a sufficient response in the other prices to influence his own demand significantly.) This situation occurs where the number of sellers in the interdependent group is *few* (or where a few control a significant portion of the total sales of the group) and where their products are significantly differentiated. This is the category of "differentiated oligopoly."
2. An industry or group of a few sellers with recognized interdependence of prices for their several products, but with slight or insignificant differentiation among these products. This is the "pure oligopoly" category, and is essentially a variant of the more general category of oligopoly with product differentiation.
3. An industry or group of sellers with differentiated but close-substitute products, where the sellers do not recognize direct interdependence or rivalry among themselves. A price change by any one seller will not elicit a significant price response from the others, but a concurrent change in all the other prices will substantially affect the demand for any one seller's output. This situation occurs where the number of sellers in the group is large, and where none of them controls much of the total market. We have previously labeled it as "monopolistic competition."

SUBCATEGORIES OF DIFFERENTIATED OLIGOPOLY

The most important market category in the American economy is that of differentiated oligopoly—the industry where a relatively few firms are rivals in selling differentiated but close-substitute products. Fewness of sellers in this context covers the situation where the total number of sellers is few (for example,

three, five, ten, or fifteen), and also may with certain qualifications be extended to cover the situation where there are fairly many sellers but where a large proportion of the output is concentrated in the hands of a few. The most significant characteristic of the market type is that there should be a degree of concentration of sales or output such that there is a close and recognized interdependence of pricing among firms which control most of the output of the industry. Oligopoly might thus in an appropriately flexible sense include such subcases as:

1. A very few sellers—for example, five sellers control the entire market, in equal or in varying proportions.
2. Heavy concentration of industry output in the hands of very few firms, with dispersion of the remainder among relatively few—for example, three sellers control 85 percent of the industry output, and 10 small firms divide the remainder in varying proportions.
3. Moderate concentration of industry output in the hands of relatively few sellers, and dispersion of the balance among quite a few firms—for example, eight firms control 70 percent of the market, and the balance is dispersed widely among 30 or 40 small firms. (This is really oligopoly with a "competitive fringe.")
4. "Quite a few" sellers of about equal size, their number being small enough that there is some recognized interdependence in pricing—for example, 20 sellers divide the market in about equal proportions.

Although these subcategories differ in significant respects, their similarity is dominant. In each case, pricing will be dominated by the recognized interdependence and rivalry, especially within the central core of large firms. At the same time, the varying degrees of concentration and dispersion may give rise to associated differences in behavior which are also significant.

Why are these varying patterns of concentration found? For the existence of concentration in general, a primary reason is the economy of large-scale production. In many industries, relatively few firms of optimum scale may be able to produce an aggregate output sufficient to supply the entire market at a

price equal to or above minimum cost. In these instances, competition has sometimes tended to drive out firms until the number was reduced to an efficient level, or, more frequently, firms have forestalled such competition, and also increased their efficiency, by combining to reduce the number and enlarge the individual size of firms. When combinations to reduce the threat of competition have proceeded a certain distance, the remaining firms may combine further for various reasons, and the number of firms may become smaller than economical production requires. This result may also ensue as each of an initial group of a few firms grows with the market through time, eventually attaining greater than optimum size. This suggests why concentration in an industry may occur, and the number of rival firms be few. But it should be emphasized that in such industries the number of firms is not necessarily stable or determinate. The dominant uncertainty in oligopoly, to be discussed below, means that the number and size distribution of firms can vary over a considerable range within many industries and be as stable at one point as at another. Such historical stability in the structure of our concentrated industries as may be observed is due principally to certain barriers to entry which often protect an established group of firms from inroads by newcomers. Most important of these barriers are:

1. The fact that a new firm of efficient size would be quite large (economies of scale being important) and that its entry would therefore saddle the industry with overcapacity and possibly engender destructive competition. Hence, the large investment required for entry is not risked.
2. The advantage in branding, sales outlets, and good will enjoyed by established firms—an advantage which makes it difficult or unduly expensive for a newcomer to build a demand for his output.
3. Legal and institutional barriers to entry set up by established firms—including full control of strategic resources or control of patents on strategic techniques. Such barriers are the most certain and effective in maintaining the status quo in an industry.

Because of such barriers to entry, the observed concentrated structure of many industries may be stable or only slowly changing through time. It should be noted, however, that the *threat of entry*, over or in spite of existing barriers, is a conditioning factor which may clearly influence the price and output policies of oligopolists.

A large number of the markets for manufactured and processed goods fall in the *differentiated oligopoly* category. The following is a partial list of familiar industries with this sort of structure: automobiles, rubber tires, gasoline and allied products, electric refrigerators, radio sets, electric razors, vacuum cleaners, soap and soap chips, cigarettes, fountain pens, prepared breakfast food, aircraft, farm machinery, distilled beverage liquors. The problem before us is what sort of price and output policies characterize industries of this kind.

PRICING WITH INDEPENDENT ACTION BY SEVERAL SELLERS

Let us look first at the logic of oligopoly pricing to see whether any definite suggestions emerge. The single oligopolist with a differentiated product—for example, the seller of a brand of electric razors, with a distinctive and protected design and other substantial or ephemeral distinguishing qualities—may be viewed for purposes of experiment as a monopolist with a distinct demand for his own product, represented very tentatively in a "demand curve," as in Figure 28. This curve suggests that at each of several prices he can currently sell each of several corresponding quantities. His situation seems to resemble that of a single-firm monopolist, but in reality it differs decidedly in two respects:

1. If the prices of all other products, and especially of close-substitute products, were given, the oligopolist would have a definite demand curve but it would be generally much more elastic than the single-firm monopolist's demand curve. This is because the oligopolist's product has close substitutes and the single-firm monopolist's product does not. If the oligopolist lowers his price, *and if other prices are unchanged*, he will gain sales rapidly as buyers shift from other products to his. Con-

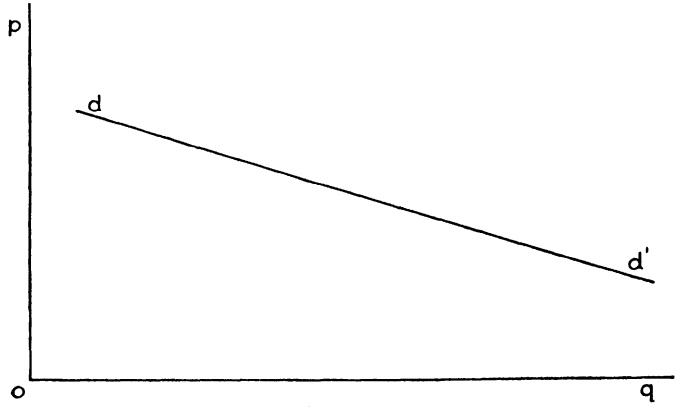

Figure 28

versely if he raises price independently, and other prices are unchanged, his sales will fall off markedly. The demand for the oligopolist's product is thus very sensitive to his *independent and unmatched* price changes.

2. But the oligopolist cannot assume (in the absence of agreements) that other sellers' prices will remain practically unchanged in the face of price changes by himself. In fact, since any gain in sales he makes by reducing prices will be at the expense of a recognized and significant loss in sales by some rival or rivals, he should probably assume otherwise. That is, each oligopolist, if he acts independently, must assume that any price change he makes will set in course retaliatory or compensatory price changes by his rivals. These reactions, which the oligopolist must expect, will affect the demand for his own product quite substantially. Unless he can guess the exact course of all rivalrous reactions, therefore, he cannot know just what effect a given independent price change of his own will ultimately have on his own demand. He can thus no longer operate on the supposition of a fixed demand curve showing the independent relation of his price to his quantity of sales, since the assumption of other prices remaining substantially unchanged is invalid. The provisional "independent" demand curve in Figure 28 must therefore be crossed out, except possibly for very small price changes.

Does the oligopolist have any way of knowing how his sales are related to his price, admitted the probability of interdependent pricing by his rivals? Not in the absence of agreement, convention, or formula governing the action of the group or of some mutually recognized and established *modus vivendi* which amounts to the same thing. This is because the course of rivalrous price reactions, and of reactions to those reactions, and so on indefinitely, could otherwise seldom be logically predicted or reasonably guessed. The independent oligopolist has no definite demand curve for his product, which shows how his independent pricing will affect his sales. What he does have at any time is his going price and his going quantity of sales, and uncertainty respecting the effect of any independent price change he may contemplate.[1]

Without the need to place cost variations against demand, therefore, a general principle can be enunciated to the effect that, *if we suppose individual sellers to act independently* (with no formal or tacit understanding or recognized convention on pricing) no certain price behavior can be attributed to oligopolistic markets. Abstract analysis assures us that the result is logically indeterminate. Something *does* happen in such markets, however, and not necessarily by pure accident. We should therefore press our general analysis a bit farther.

Businessmen are not particularly fond of uncertainty. If placed at the mercy of chance, they will attempt to escape this position through some remedial action. It is therefore legitimate to speculate about their ways and means of dealing with oligopolistic uncertainty, drawing so far as possible on observations of how they have acted in such situations.

Uncertainty in oligopoly pricing becomes dynamic only if some oligopolist changes his price. If everyone adheres to a

[1] Some writers have suggested that even with "strict independence" in pricing by several sellers, mutual recognition of interdependence *could* cause each to set price and respond to others' price changes in a determinate fashion designed to maximize combined profits. The assumptions subject to which this is true are unrealistic, and in any event may be taken as involving the assumption that each seller in turn has decided to accord his rivals a position in and definite share of the market, and knows that his rivals have decided the same. Such a situation is better recognized as one of tacit understanding among sellers than one of "independent" pricing.

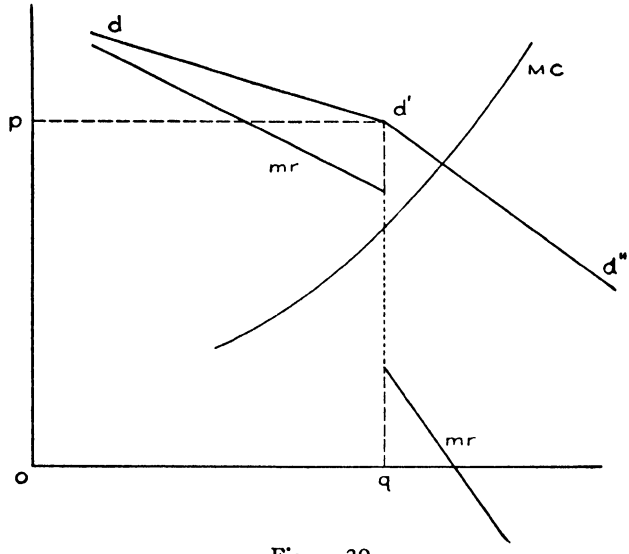

Figure 29

going price, then each may rely on past experience to predict some relatively dependable sales volume. The uncertainty in oligopoly, especially if the oligopolists make relatively independent decisions, thus tends to discourage large or frequent price changes. Policies of *price rigidity* are strongly encouraged. Any price position at which independently acting oligopolists arrive which is relatively satisfactory to profits tends to be perpetuated in the face of minor and even major changes in cost and demand.

This tendency (or one instance of it) has been described diagrammatically in Figure 29. At any moment the independent oligopolist with a differentiated product sells some quantity q at some unexplained price p. If he raises his price above p, he may assume that his rivals *will not* change their prices, and that his sales will fall off rapidly along the elastic line dd'. If he reduces his price below p he may assume that his rivals will match his reduction (or worse) in which case his sales will increase slowly along the less elastic line $d'd''$ (which has the same general elasticity as the "market demand" for the whole industry). Hence the seller's demand curve is a bent line which

turns at the going price *p*. Correspondingly his marginal receipts are discontinuous—*i.e.*, they are positive and high at prices above *p*, and they are low or negative at prices below *p*. Hence it may be that *MC* is less than *mr* above *p*, and *MC* is greater than *mr* below *p*. So the firm will "maximize its profit" by retaining the arbitrary price *p*, provided that at this price the relation of price to average cost allows what is regarded as a "fair" or satisfactory profit.[2] In this event, the prices maintained by the several oligopolists are not necessarily such as to maximize their combined profits.

This is a possible pattern. But it should be noted that the explanation tells us only why price may remain where it is. It does not tell us how it got there, or how high it will be, and infers that this may be a matter of accident. It tells little about how large oligopoly profits are or how average or marginal costs are related to price. With independent action, the resting place for the rigid price is still uncertain. For this reason genuinely independent action by oligopolists is likely to prove unsatisfactory to all of them. And for the same reason, the "kinked demand curve" theory explains much less than it at first seems to.

COLLUSIVE AND CONCURRENT PRICING IN DIFFERENTIATED OLIGOPOLY

Oligopolists therefore often tend to seek some sort of convention or agreement with the aid of which they may move *concurrently* to a price at which they earn maximum, or at any rate satisfactory, profits. Interdependent sellers in oligopoly will ordinarily seek: (1) a means of initially arriving at a satisfactory general level of prices for all of them—ideally one such that the aggregate profits they share are a maximum; and (2) a means of securing concurrent changes in price whenever changes in demand and cost conditions indicate that this would enhance profits. The several sellers of a differentiated good, like cigarettes, thus have an incentive to act together as a single

[2] See Paul M. Sweezy, "Demand Under Conditions of Oligopoly," *Journal of Political Economy*, vol. 147, pp. 568 ff.

monopolist, viewing the total market demand for the sort of good they produce as a common property to be exploited in common. This does not mean that there is no tendency to rivalry among them, for that may remain quite strong. But the unstable and unpredictable results associated with truly independent pricing force them toward concurrent action in setting price.

The technique adopted to secure concurrent pricing action may vary from the formal agreement to the extremely informal understanding or convention. Some of the principal alternatives found in practice are the following:

1. The "full cartel" or formal agreement, whereby the rival sellers contract among themselves to set uniform or related prices and often to set output quotas for each seller, together possibly with exclusive market territories, limitations on capacity, etc. These cartels are in a sense a logical solution for oligopolies, and in many countries they are common. Cartels are generally illegal under the American antitrust laws, however, and as a consequence those we have in the United States (except for a few set up under patent licensing privileges) must operate in secrecy and are of course not enforceable at law. As a consequence, American firms in concentrated industries more frequently adopt alternative means of securing concurrent pricing action.
2. The *subrosa* collusive agreement covering selling price or division of the market, but leaving open competition on matters not covered; *e.g.*, a price agreement leaving open rivalry for sales volume via product development, advertising, and so forth. These are also illegal in the United States, but in a good many industries they have operated from time to time at least.
3. The convention of *price leadership*, whereby the various sellers in an oligopoly recognize one seller (often the largest) as a "price leader" and follow his price changes closely. Rivalry of a nonprice nature, whereby the individual seller seeks to enhance his share of the market, is ordinarily left open. This type of convention is legal in the

United States so long as no collusive agreement can be found. It is one of the most common types of pricing convention in all sorts of oligopolies in American industry.

4. The mutual adoption of certain common formulas for computing price, which when used by all sellers will result in identical or closely similar prices. The most familiar formula involves the addition of a customary mark-up percentage to the average total or average variable cost per unit of output, as computed according to certain rules or conventions. Where the several sellers have also adopted a uniform cost-accounting system, this type of action can result in identical prices. Such activity is legal in the United States if no collusion is demonstrable. (Where "independent" pricing by sellers with mutually recognized interdependence gives fairly determinate results, it must generally be by tacit elimination of full independence in favor of some such arrangement as the two preceding.)
5. Finally, some concurrence in price action may be secured if the rival oligopolists simply all follow conservative price policies (each having ascertained by past experience that he may expect his rivals to follow the same convention)—if they change their prices only occasionally and in response to major shifts in costs, and follow the habit of just matching their rivals' price changes.[3]

[3] In logic and from observation it is evident that the loose concurrence secured through "mutual respect" works well only where the differentiation of the various sellers' products is substantial enough that the various sellers need not have identical prices in order to maintain some balance in the proportions of the market they hold. Thus various automobiles differ sufficiently (and in a sufficiently unmeasurable fashion) in design, size, and quality, that constant identity of Ford, Chevrolet, and Plymouth prices is not necessary to market stability. Nor is there any necessary steady differential between the Pontiac (General Motors low-medium price class) price and the Chrysler (Chrysler Corporation high-medium price class) price. In such a situation, loose concurrence in pricing is all that is required to avoid chronic instability. Where the product differentiation is slight, however, constant identity of rival sellers' prices is a condition of market stability. In the gasoline market, for example, a half-cent-per-gallon differential between any two major brands would seriously weaken any "brand monopoly" of the higher-priced seller. In such cases either price leadership or collusion will ordinarily be resorted to.

One or another of the means of securing concurrent price is common in most oligopolistic industries, including differentiated oligopolies.[1] The degree to which rivals secure concurrence in pricing, the way in which price is related to cost, and the manner in which price responds to changes in cost and demand through time will vary somewhat with the type of cooperative action in use in the industry, and with the area of unrestricted rivalry left open.

Where there is effective collusion on price—as with a full cartel, collusive agreement, or price leadership—so that the firms or their leader are free to choose for the industry a "best possible" price in any situation, some central tendency is probably predictable.

As a first approximation only, and for the moment overlooking selling costs, such concurrent oligopoly pricing tends to the single-firm monopoly level. If the oligopolists concur and act together on price, they should tend to choose a price in the neighborhood of that for which their combined marginal costs equal the marginal receipts from aggregate sales. Their calculation begins with an estimate of the industry demand, at each price the sellers may jointly adopt, for the sort of good they produce. To the resulting "demand curve" may be drawn a marginal receipts curve. The oligopolists can then maximize the aggregate profit which they share by choosing the common price and combined output at which their aggregated marginal costs equal these marginal receipts.

It may thus appear that the collusive oligopoly price tends to be very much like the single-firm monopoly price. Several important qualifications to this conclusion, however, must immediately be inserted. First, the combined demand for several substitute differentiated products is a much less certain conception than the demand for a single good and involves much guesswork and room for disagreement among the several sellers. Similarly, different oligopolists may have different marginal and average costs, and therefore different ideas of the profit-maximizing price. Further, the choice of a precise "monopoly"

[1] See A. R. Burns, *The Decline of Competition* (New York, McGraw-Hill Book Company, 1936), for discussions of commonly employed methods of price making.

price for which the aggregate marginal costs of the several oligopolists equal the marginal receipts drawn from the market demand curve for their combined products would imply a concurrently determined sharing of the market by the several sellers, with shares such that the marginal cost of production for all sellers would be equal. Collusion is ordinarily not complete enough to establish such market shares, but often leaves these open to determination by nonprice competition, thus perhaps widening the area of disagreement over the most profitable price. If shares were fixed, moreover, as in a cartel, they would probably not recognize the marginal cost principle, but would depend on the relative bargaining strength of the rival sellers. The general "monopoly price" at which oligopolists might arrive, whether by explicit agreement or by price leadership, would therefore not be a precise point, but would lie at some indeterminate point within a significantly wide range of prices. It may often, moreover, be arrived at without deciding exactly what the outputs of various individual sellers will be. This is mainly because collusion is generally imperfect, as a result of differences in the cost conditions of sellers, of differences in their attitudes, of their several desires to remain somewhat independent, and of the threat of the antitrust laws.

Second, the oligopoly price is likely to be adjusted less frequently than a single-firm monopoly price in response to shifting demand and cost. As demand and cost curves shift back and forth with cyclical business fluctuations, an oligopoly is less likely to make short-run adjustments to the changing situation than is a single-firm monopoly, and much less likely to do so than a purely competitive industry. This is in part because when price is controlled by agreement or price leadership, every change in price places some strain on the controlling mechanism and increases the probability of defections from the agreement. By maintaining a relatively rigid price, the oligopolists lessen the possibility that open price rivalry or price cutting will emerge. Oligopolists may therefore try to set a price which will be workable for a wide range of demand conditions.

When prices are set on this basis, any very close matching of short-run marginal costs and marginal receipts is obviously not attempted. The tendency of oligopolistic firms to pursue rela-

tively inflexible price policies is enhanced if their short-run average variable and marginal costs of production are relatively constant over wide ranges of output. Extension of output with given wage rates and raw material prices will thus be possible without much rise in out-of-pocket cost, and with declining average total cost, up to rather high levels of output. If rising wage rates do not cause costs to shift upward too rapidly the firms may find a rigid price consistent with long-run and not too inconsistent with month-to-month profit maximization. But the long-run price may also vary in a significant range around that at which long-run marginal costs would equal the marginal receipts from the long-run average demand.

Third, price leadership or agreements will occasionally become ineffective under the strain of depressed business conditions, or perhaps because of disagreements among managements of rival firms over labor policy or other matters. Monopoly pricing will then give way to price warfare with very low prices for some interval.

Fourth, the effect of *potential new entry* to the industry may have a very significant effect—much stronger in all probability than in single-firm monopoly. We have already indicated that the single-firm monopolist (p. 156, Fig. 27), may choose not to go above a certain price, even though it would be immediately profitable to do so, because by so doing he would induce entry and face competition and lower profits. Therefore if he can forestall entry at some limit price and still anticipate greater long-run profits than if he attracted entry, he will charge such a price and continue to enjoy his monopoly. In this event, his demand curve is truncated at the limit price, and his marginal cost need bear no unique relation to price or to marginal receipts. The same reasoning may be applied to oligopolistic price policies. Should any number of established oligopolists recognize the threat of further entry and reason that it would be more profitable to stay at or below some limit price at which entry could be excluded than to charge more and attract entry, they may also choose to hold to the limit price. And in this event it is quite possible that their outputs may be such that marginal cost could exceed rather than fall short of price.

This result is not different from but consistent with monopolistic pricing where the monopoly is subject to a similar threat of entry. But since the threat of entry is likely to be much more common and stronger in oligopoly than in single-firm monopoly, this type of reasoning is likely to play a much stronger role in the oligopoly case. Where it does, as we shall see below, the price the industry charges, although potentially consistent with long-run profit maximization, may be arrived at without direct reference to marginal cost and receipts, but may instead be made by adding a "safe" (*i.e.*, entryproof) percentage mark-up to average total cost. And in this case, not only is the price–marginal-cost relation for any firm unpredictable, but the relation of price to average cost and the size of industry output depend mainly on the cost and other advantages which established firms enjoy over potential entrants. The general level for a rigid price, suggested above, may frequently be set just so as to permit the established sellers to exploit their position as profitably as possible without attracting additional entry.

It is, of course, possible that the established sellers in an oligopoly either will neglect the possibility of entry, or, recognizing it, find that a price to exclude it would be less profitable in the long-run than a higher price which would attract it. In either event, they may then exploit the immediate possibilities of their industry demand curve, taking no account of induced entry. Where they do this, and thus set prices high enough to attract new sellers into the industry, entry may so reduce the demands of individual sellers that most or all of them can make only normal profits even at the most profitable price. In this instance, an initially high price plus ease of entry may eventually result in excess capacity and uneconomically small-scale plants in the industry.

Suppose, for example, that each of six rival oligopolists finds himself in the position shown in Figure 30, where *dd'* represents each seller's *share* of the industry demand *on the assumption of uniformly concurrent pricing*. Each seller might currently exploit his position at the price *op*, where each would make the excess profit *abcp*. But this excess profit might induce several more sellers to enter the industry, thus dividing the

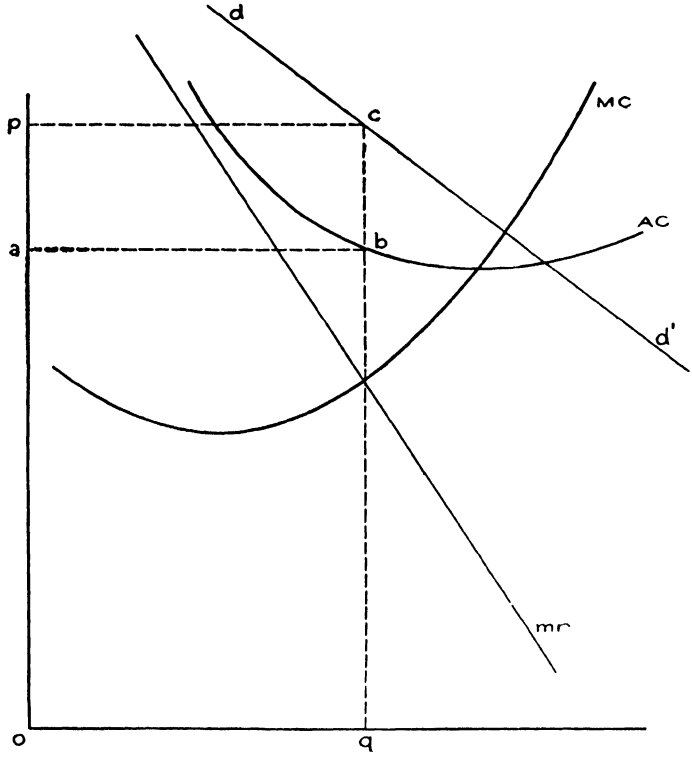

Figure 30

total industry demand into more and smaller individual shares. Each seller's share of the market demand (*dd'*) would shift leftward. This adjustment *could* bring each seller to the position in Figure 31, where price is still at a high monopoly level but now equals average cost. Excess capacity and small scale have driven average costs up to the level of a monopoly price. Identical results for all sellers are of course improbable, and the whole process would depend on the original oligopolists' failure or indisposition to discourage excessive entry through a lower price policy. But such a negligent policy and an outcome seem distinctly less probable, and also appear in fact much less frequently, than the alternative of forestalled entry at a limit price suggested just before.

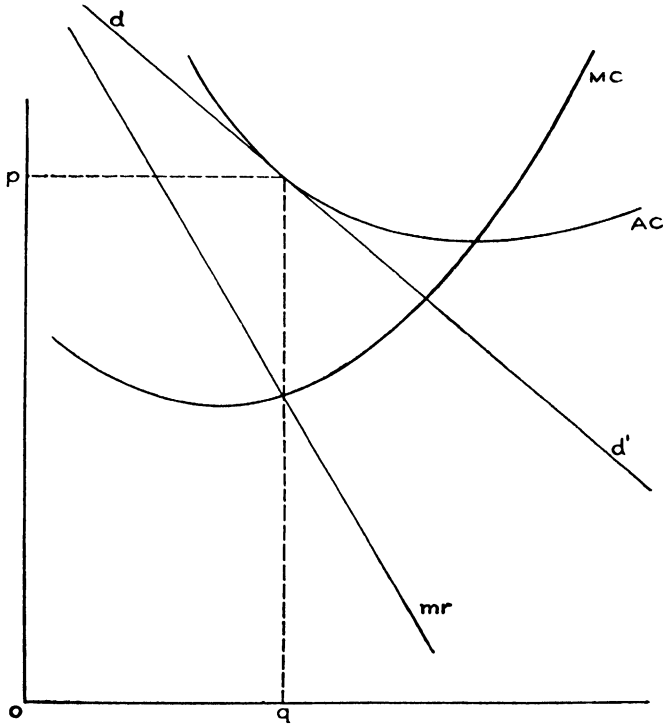

Figure 31

In the various possible oligopoly equilibria so far discussed, two main things are apparent. First, the relation of price to marginal cost can be extremely variable from one oligopoly to another, although there is some general tendency toward outputs such that marginal cost is less than price. Moreover, the marginal costs in question may frequently arise from situations where the aggregate costs of successive industry outputs are greater than the minimum attainable aggregate costs of these outputs. (In this respect also oligopoly is similar to monopoly.) Second, the relation of attained aggregate cost of industry output to minimum attainable aggregate cost of this output may also be extremely variable. With a chronic excess in the number of firms, the discrepancy may be serious; in other cases, it may not. It is possible under different conditions that the number of firms

might be too large, just right, or too small from the standpoint of economical production. As in monopoly, the attainment of minimum aggregate costs for any given or ideal industry output may not permit the firms to operate at optimum scale, although as the number of firms appropriate to the industry increases, closer approximations to this result should be feasible. But at least a part of the oligopolists' departure from optimum-scale operations will ordinarily be socially undesirable, as caused by either excessive or insufficient entry, or by undesirable restriction of industry output.[5]

So far we have been speaking of probable oligopoly behavior where there is a cartel, or a collusive agreement on price, or an effective convention of price leadership, whereby the sellers choose jointly or by a trusted representative some "most profitable" price. For even this relatively simple case, it is clear that the price or prices which oligopolists pursuing a maximum profit may set may be in very loose approximation to any uniquely predictable monopoly price. Differences in the attitudes and positions of rival sellers set up a range of dispersion within which price may fall; the mechanics of collusion suggest a more rigid or less frequently adjusted price; and price making may be dominated by the threat of entry, so that average costs plus a "safe" mark-up effectively determine price, and price–marginal-cost relations are substantially unpredictable. Subject to these limitations, a collusive oligopoly price and output may be characterized as similar to those of a single-firm monopoly.

The approximation is not necessarily poorer if the means of securing concurrent pricing is even looser—as where sellers in an industry adhere to a formula for price making by adding a customary mark-up percentage to average total costs of production. It is not uncommon for sellers in oligopolistic industries—especially in those not highly enough concentrated to allow very close collaboration—to make prices by computing the "normal" or "long-run" average cost per unit of product and adding 8, 10, 15, or some other percent considered usual or proper in the industry in order to arrive at price. The "normal" cost is ordinarily computed with fixed cost allocated in such a way that

[5] Cf. pages 153-154.

average fixed cost is not much influenced by short-run variations in output, and corresponding variations in average variable cost may similarly be subdued. In effect, price may be based on a sort of long-run average cost for long-run average output and may be influenced strongly only by variations in wages and other factor prices.[6] A similar pricing procedure is found where merchants make price by adding a customary percentage to the direct cost of the goods on the shelves. Now it will be noted that this sort of pricing policy is potentially quite consistent with one of (1) avoiding frequent price changes and adhering to an all-purpose rigid price for considerable periods, and (2) making price at some limit level designed to forestall additional entry. It is quite possible that in setting prices by such seemingly arbitrary mark-up formulas, sellers are moving directly to the rational profit-maximizing goal—the most profitable price which can be had without bringing in too many competitors. At the same time, the relation of price to marginal cost is not always predictable in this instance, and the approximation to "ideal" monopoly pricing may not be very close either on the average or at any one time.

Many sellers in concentrated industries nevertheless follow such a pricing policy and by so doing succeed in making roughly concurrent pricing decisions and in avoiding price rivalry. They evidently do so because collusion is difficult to arrange, legally hazardous, or unwelcome to some sellers; because their estimates of demand conditions would be so highly uncertain that they prefer a quasi-arbitrary formula to a hazardous calculation of marginal receipts; or because their thinking is dominated by the threat of additional entry. So long as the basic average costs are calculated so as to neglect short-run variations in average fixed cost, however—so that prices are in effect based on short-run average variable and perhaps[7] on marginal costs—and so long as the mark-up percentage is rationally adjusted to the elasticity of industry demand or to the threat of entry, then there is nevertheless some meaningful approximation to monopolistic price and output policy.

[6] See R. L. Hall and C. J. Hitch, *Price Theory and Business Behavior*, Oxford Economic Papers No. 2 (Oxford: Oxford University Press, 1939).

[7] If average variable costs are constant.

Where the sellers in an oligopolistic industry simply follow conservative price policies—adding ample but perhaps variable mark-ups to the average cost of output and taking some pains to match each other's prices if their several calculations lead to different results—the price and output result may well be of the same general order as that already described. Price may not be made directly by independent marginal calculations, but, as there is "more than one way to skin a cat," the outcome may be a meaningful approximation—subject of course to a significant range of error—to monopolistic pricing. But in this connection monopolistic pricing must be taken to include "limit pricing" which holds current price down to forestall future entry.

In view of all of the preceding, may anything be said of the long-run average resting place of the price–average-cost relation and the price–marginal-cost relation for the individual firm in the oligopolistic industry? The price–average-cost relation, and the size of excess profits, will evidently vary among industries, particularly with the degree of ease of entry and with the way in which established sellers react to the threat of additional entry. Thus the position shown in Figure 30 may show the central tendency of the range within which price may fall for the firm in tightly closed oligopolies with effective price leadership or collusion and no threat of additional entry. Here the individual seller succeeds in exploiting his share of the industry demand (dd') fully through concurrent pricing with his rivals, earns an excess profit and, of course, has marginal cost equal to marginal receipts and below price.

The position shown in Figure 31 might characterize the firm, in oligopolies where entry was easy, and was attracted because established firms attempted to maximize their profits for the short run by concurrent pricing. Here entry has been induced to the point where each firm (of the less fortunate ones) has so small a share of the market (dd') that it can just break even at the monopoly price, which here equals average cost and exceeds marginal cost. The position in Figure 32 might characterize the seller in an oligopoly where the several sellers were following a relatively low limit-price policy to forestall entry. Here the individual seller forsakes exploitation of his share of the industry demand above the price p, since this would induce

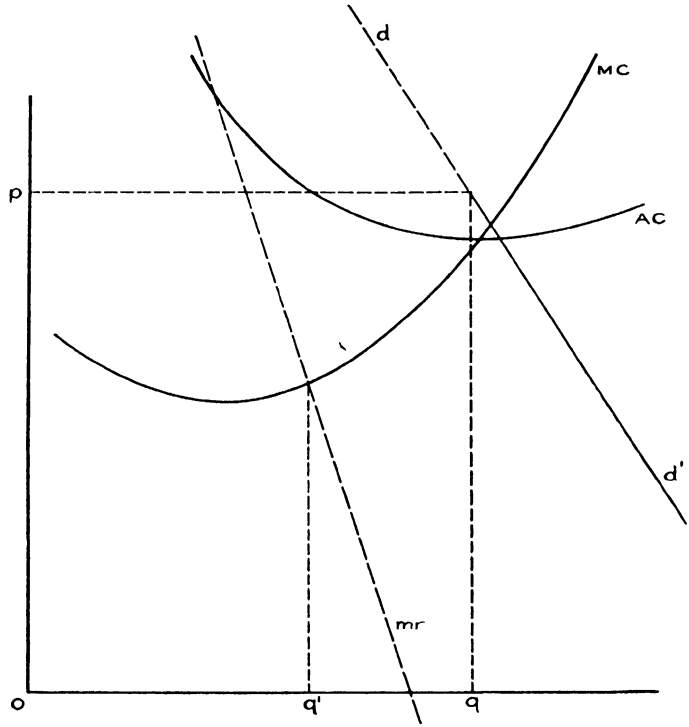

Figure 32

entry. As a consequence he does not produce the output q' for which his MC equals mr. Instead he charges the limit price p at which it turns out that he sells the amount q. In the illustration, price exceeds average cost (though not by enough to maximize short-run profits) and also exceeds marginal cost slightly. But it could as well be equal to or less than [8] marginal cost, and the price–average-cost relation could vary from industry to industry, depending on the threat of entry.[9] The limit price p

[8] Here the seller furnishes some units at a price less than their current marginal cost in order to fulfill market demand at the limit price and thus keep price low enough to exclude undesirable entry. Such marginal losses might be regularly accepted to protect long-term gains.

[9] And the aggregate cost of any attained output bears a potentially variable and not uniquely predictable relation to the minimum attainable aggregate cost

might be arrived at either through direct calculation of the threat of entry, or through some conventional mark-up formula. It will be noted that in the third (and perhaps most usual) case, there is a probability of some excess profit but no generally predictable relation of marginal cost to price.

MARKET SHARES IN DIFFERENTIATED OLIGOPOLY

This gives us certain alternative conclusions, overlooking selling costs, concerning the general level of an oligopoly price chosen through concurrence of rival sellers. It does not, however, indicate how the various oligopolists may share the market which exists at the prices they establish. This is an important consideration, for within a general agreement on price there may be a very active rivalry for sales volume carried on either by "secret" pricing tactics or by *nonprice competition.*

The aggregate amount sold by the several oligopolists (and the aggregate profits they earn) will depend on the general level of price at which they arrive, by concurrent action or otherwise. This price may be at a "monopoly" level, which fully exploits the aggregate market demand for the goods in question, or it may be at any lower level. Its level depends upon how much unanimity of opinion on price policy the oligopolists can secure, how effective a collusion they can maintain, what the threat of additional entry may be, and how they react to this threat. Experience suggests that some oligopolies have been more successful than others in securing harmony on pricing and in maintaining high and exceptionally profitable prices.

Regardless of the general level of price secured, however, the oligopolists must still share the total market volume. Suppose that major cigarette manufacturers have arrived at a general concord on a retail price for their main brands of 14 cents per package. What determines the relative shares of the market secured by Camel, Chesterfield, Lucky Strike, Philip Morris, and Old Gold? And also, what determines the share secured

of that output. Further, product differentiation undermines somewhat the precision of the concept of an industry's marginal cost without furnishing a very meaningful substitute.

by several other cigarette manufacturers who are in the market with, let us say, Zeros, Ropos, Hempos, and other brands?

The strategic consideration in this regard is that the several substitute products are differentiated—are viewed by consumers, correctly or incorrectly, as somewhat different products. If the products were absolutely undistinguished and undistinguishable, their prices would have to be identical. Any seller with a higher price than others would sell nothing. Since the products are differentiated, however, their prices can be somewhat different, especially if there are real or supposed differences in quality. The relative proportions of the market secured by various sellers will be fairly determined by (1) buyer preferences and (2) price differences. Thus, in the cigarette market, the shares secured at 14 cents per package by Camel, Chesterfield, Lucky Strike, Philip Morris, and Old Gold are determined at any moment by the state of buyer preferences among them, prices being equal, and by the relative preferences of buyers between these 14-cent cigarettes and others of different quality sold at 18 cents, 16 cents, 12 cents, and other prices. Assuming that the oligopolists have arrived at some concord on the general level of prices, then, their relative shares will be determined both by the price differentials they establish and by the success they have had in designing and advertising their products to secure buyer preferences.

Establishment of stable price differences which are generally announced and known is not always possible, since differentials may be the source of recurrent instability in the general level of price. When rival oligopolists have announced price differences, these are generally based on the supposition by buyers that there are corresponding differences in quality of product, since otherwise the higher-priced sellers could not allow the price differences to remain. If such price differences become a regular part of the industry pattern, there must be struck, by luck or by experiment, a nice balance between price differential and quality differential as evaluated by buyers. Ordinarily the higher-priced sellers must have been able to build up a considerable product differentiation in their favor. When they have been able to do so, a complex pattern of price and quality dif-

ferentials may prove relatively tenable and not be conducive to general price instability or warfare.

There is always the chance, however, that a price differential may degenerate into a price cut, being set so that lower-priced sellers begin to "take business away" from higher-priced sellers. In a number of industries where product differentiation is not too great, this potential difficulty has become actual, with the result that retaliatory price-cutting rivalry has tended to emerge periodically.

The gasoline market of the petroleum industry furnishes a good example of this. Small sellers of lesser-known brands of gasoline stay in the market by charging one or two cents less per gallon than the larger sellers of well-known brands. But the discovery of a satisfactory differential at which all sellers can maintain over time a fairly stable share of the market has proved very difficult. As a result, the large sellers have periodically begun price-cutting wars to regain their market positions. We may therefore add to our catalogue of generalities about differentiated oligopoly pricing that the difficulties of finding appropriate price differentials may in some cases make concords on the general level of price inherently unstable. The level of price may then be intermittently reduced as the result of price warfare. A concomitant of this tendency is that sellers in such markets tend to seek collusive agreements which stabilize interseller price differentials, even though such agreements are generally contrary to law.

The price differences which assist in determining various sellers' shares of the market can be established, with less chance of stirring up discord, by special, secret, or unannounced price reductions. Where the market is such that flexible price concessions can be readily made, individual sellers can ordinarily adjust their net prices so as to maintain a given position in the market without at the same time setting in course retaliatory price cutting. Secret and special price concessions are frequently possible in markets for producer's goods. They can be made by allowing discounts from the announced price, on the basis of quantity of purchase or type of buyer, or simply by negotiating special terms for big orders.

One consumer good which is ordinarily priced at retail on a flexible basis is the automobile, which is sold "to meet local competition" at varying net retail prices by varying the allowance on the used-car trade-in. Where several oligopolists follow the policy of announcing given prices for their goods, but of allowing price concessions of various sorts, they aim at maintaining a favorable general level of price, and at solving the apportionment of the market among them through flexible deviations from this price. Certain industries have been quite successful in pursuing this policy. But in times of declining income, when the total market is being severely reduced, individual price concessions have often degenerated into general price cuts and price wars have resulted. Reacting to this threat, a good many oligopolistic industries attempt to limit or standardize discounts and concessions by agreement or by trade-association activity, thereby enhancing the prospect for a stable price.

In sum, the general oligopoly tendency to concur on a favorable price level is tempered by the tendency, especially in bad times, of interseller price differences to lead to breakdown of the concord and to price warring. This tendency is in turn often countered by agreements among the oligopolists regulating price differentials, discount policies, and even shares of the market, but those agreements are made uncertain by the fact that they are ordinarily contrary to the antitrust laws. What happens in the net in any oligopoly is affected strongly by chance, or by a good many small considerations which have never been thoroughly investigated. Observation suggests, however, that most differentiated oligopolies manage to maintain relatively high and stable prices most of the time, and suffer from price instability and price wars only in times of temporarily depressed demand or of rapid secular contraction of demand.

NONPRICE COMPETITION

The seller in differentiated oligopoly ordinarily wants three things: (1) a profitable and dependable general level of price for his industry; (2) a stable or a growing share of the market for himself; and (3) a means of obtaining such a share without precipitating price rivalry and price instability. In devising price

and market policies, therefore, most such sellers attempt to avoid consistent employment of price changes or price competition as a mode of rivalry, and to emphasize nonprice rivalry in the form of product variation, advertising, and other sales promotion. (This, of course, if full cartelization has not eliminated all sorts of competition.) They tend to accept a price at which general concord with rivals can be maintained, and then to vie for a stable or growing share of the total market by incurring various product-development and selling costs.

This generalization is based simply on observation of a large number of industries, like those producing automobiles, cigarettes, electric refrigerators, and fountain pens, but it has a fairly sound logical explanation. Nonprice rivalry seems less likely to degenerate into unbridled warfare than does price rivalry. Though rival oligopolists may run up large competitive selling outlays, they probably find it easier to keep this sort of rivalry within profitable bounds. This may be because of the greater institutional frictions encountered in extending selling cost, or because a rival's selling or product policy may be matched by ingenuity as well as by gross money outlay.

Most sellers also probably feel that a given amount spent on skillful product variation or sales promotion is a "better gamble" to gain sales volume than an *equivalent* concession in the price of the product. Thus a cigarette manufacturer selling 500 million packages of cigarettes per year is likely to prefer initiating an expenditure in addition of one-half cent per package, or 2½ million dollars, on popular radio programs and periodical advertising, to initiating a one-half-cent reduction in the price of his cigarettes, assuming that either move may be matched by his rivals. Similarly, an automobile manufacture going into a new year is very likely to continue to charge $1000 for his lowest priced model rather than $950 and to put the $50 difference per car into the adoption of new body designs or mechanical features. To be sure, advertising or product variation will be matched by rivals as surely as price cuts, but less easily, less quickly, and less certainly, thus giving the individual seller a better chance to gain an edge through his own ingenuity. Preference for nonprice competition may also stem from the belief that the buyers of consumer's goods (the usual output of differ-

entiated oligopoly) are more product-conscious and advertising-conscious than they are price-conscious.

The fact is, in any event, that sellers in differentiated oligopoly do compete for larger shares of their markets principally by product variation and sales promotion. As a result they incur substantial product-development costs and selling costs. It is not possible to determine logically the point to which such costs will be carried in oligopoly. The relation of selling cost to production cost or to price is indeterminate. But observation suggests (1) that relatively large selling costs (including costs of product variation), running frequently as high as 20 to 30 percent of total cost, are incurred; but (2) that such costs are not usually pushed to the point where excess profits are entirely eliminated. In effect, some restraint in the extension of selling costs is ordinarily observed.

The relation of the individual seller's costs to price is far from precisely determinate in the situation described. It may not be unusual, however, for his situation to be as shown in Figure 32. Suppose there that *AC* shows his average total costs, *including a given outlay on selling costs*, and *op* shows his chosen price, as influenced by whatever price concord he has with his rivals. At this price, and with chosen selling outlays, he is succeeding in selling some output *oq*. He has extended selling cost as far as seems currently feasible in view of rivals' possible reactions to further sales promotion on his part. He will not cut price for fear of inducing retaliatory price cuts. The line *dd'* shows how much his sales would change if he changed his price and if rivals matched his actions. Accordingly, generally lower prices would be less profitable to all, and are avoided. Higher prices would be immediately more profitable but would induce new entry or otherwise disturb the market, and are therefore also avoided. The individual seller has reached no "profit-maximizing" equilibrium (which would equate marginal cost and marginal receipts) but is striving to hold a profitable share of the market at a favorable price, without upsetting the going price and without so far extending his selling costs as to dissipate all his returns in this way.

Price, however, may be higher or lower than that shown in Figure 32, and output (relative to the optimum) smaller or

larger, with varying effects on profits. It is also possible that price will initially be set high enough to attract excessive entry, leading eventually to a high-price, low-profit position as shown in Figure 31, or that competitive selling outlays will drive average costs up sufficiently to produce the same result. It is furthermore possible that concurrence will extend to placing some limitation on selling outlays, thus protecting favorable profits in this way also. Observation suggests, however, that there is usually at least a moderate amount of nonprice rivalry in differentiated oligopoly.

Generally, the various possible price-cost relations which may emerge in oligopoly with nonprice competition thus correspond to those already illustrated in Figures 30, 31, and 32, and described in preceding pages. The main significant addition is that costs, as represented in the level of cost curves will be larger because of selling and product variation costs. The cost curves will be at a higher level, and outputs may very well be virtually smaller and prices higher in order that these costs may be covered.

The result for the economy of this sort of oligopoly price policy is quite evident. Prices in the long run tend to be high enough to cover production plus selling costs and still to yield a normal or superior profit. They therefore may be higher than they would be if large selling costs were not incurred. A significant proportion of productive resources is correspondingly diverted from activity in producing to activity in sales promotion and product variation. As the cost of selling becomes higher, the immediate productivity of the economy tends to be reduced. Finally, the economy may or may not be compensated for these higher real costs through corresponding gains in quality, design, and variety of products, and through other less tangible sources of enjoyment.

An important question thus suggested is what sort of balance may be struck between the additions to cost and price which arise from nonprice competition and the associated advantages of product improvement and variety. Abstract logic will give us no solution to this problem, for there is no way of telling a priori how much consumers will benefit from a given expenditure on nonprice competition. Some light may be cast on

the issue, however, by considering the results which have been secured in familiar industries. So far as product variation and product development are concerned, there seem to be three principal sorts of results of competitive expenditure along this line:

1. The product may be improved in a substantial and tangible fashion, in either design or quality, so that most buyers feel fully compensated for the additions to cost and price. At least part of the product changes in the automobile industry from 1900 to date seem to fall in this category. Better and more efficient design has been progressively developed, to the very substantial advantage of automobile users.
2. The product may be improved in quality, changed in design, or sold with an accompaniment of auxiliary service, in a manner which is attractive to buyers but which does not necessarily give them products enough better or more useful to compensate them (according to their own standards) for the added cost. Many minor variations or model changes in the automobile industry seem to have fallen in this category. Parallels may be found in other industries producing durable consumer's goods.
3. The product may be subjected simply to nonprogressive or slightly progressive variation or change, on the order of periodic style changes, in order to stimulate buyer interest. Periodic variations in fountain-pen design, in the length of cigarettes, or in the "streamlining" of immobile household appliances are frequently representative of this tendency.

The reward which the consumer gets for the added cost of product variation thus evidently depends on the sort of product variation accomplished. Sellers adopt variations because they promise to be profitable. If they also result in much better products, all is well. Instances of all sorts can be cited. Since there is a certain admixture of useless and useful, however, we can say that for the whole range of differentiated oligopolies (including a large proportion of consumer's goods industries) the cost of progress in products is relatively high. On the other

hand, the *gross rate* of product development would probably be less under any other system. The sort of nonprice competition described will also produce within any industry a great *variety* of substitute brands, designs, and qualities among which buyers may choose, and this variety may constitute an advantage.

What can be said of selling costs devoted not to product variation but to advertising and sales promotion? Competitive advertising in oligopoly is well illustrated by the radio advertising of the cigarette industry, the soap industry, or the makers of prepared breakfast foods. Another variety of sales promotion is found in the gasoline industry, where rival sellers vie in constructing more, larger, and fancier service stations in order to induce buyers to use their several gasolines. In considerable part, the money spent on and resources devoted to competitive advertising add little or nothing to the output of the economy or the welfare of buyers. Goods become more expensive without being necessarily better or more enjoyable. Thus the jazz band aired by one cigarette manufacturer offsets the quiz program presented by another, and both simply maintain their relative positions in the market. The magazine claims of another company that its product is "less" irritating counter those of another that its product is "better" tasting, all at the expense of a good deal of paper, ink, and advertising-agency time. But the buyers of cigarettes probably do not obtain much more enjoyment from smoking because of this advertising, any more than the housewife's "washes" are whiter because of the plethora of "soap operas." The real cost of the good is ordinarily increased, but the buyer receives little more in product.

The main advantages to the economy of advertising apparently lie in its support of radio entertainment and in its sharing of the costs of publishing periodicals. Although the quality of the entertainment which results is often justly criticized, and although the effect of advertising subsidies on journals of opinion may not be wholly desirable, expenditures on advertising find their principal specific justification in the indirect or secondary benefits of paying the cost of entertainment or education.

Sales promotion of the sort found in gasoline distribution may occasionally offer more direct benefits. Thus the provision,

because of competitive service-station building, of a considerable excess capacity of service stations and of elaborate free service on expensive sites, although it does increase the cost of gasoline noticeably, provides motorists with greater convenience and service. Here again, however, the expenditure on selling is often excessive from the standpoint of return in added service to buyers.

One justification frequently advanced for large advertising expenditures is that they "enable" firms to grow to large scale by securing a large market, and thus to secure the economies of large-scale production. In this event, it is argued, the economies or saving of large-scale production offset the expense of advertising, and there is no net social loss. It is quite true that this *could* be the case. But it should be noted that if it were, it would be necessary (1) that *without* large-scale advertising, firms would be kept unduly small by the fact that there were too many of them and that each had a small share of the market—*i.e.*, there would have to be a substantial excess of entry—and (2) that when large-scale advertising was undertaken, some firms would be substantially more successful with it than others, and thus grow at their expense, so that the promotional effort would not be self-canceling. It may be difficult to identify cases where *both* of these conditions would be fulfilled.

Nonprice competition in oligopolistic industries as a whole seems to be such that (1) selling costs are a significant portion of total costs, (2) prices are high enough to cover these costs, and (3) the increment to buyers' welfare is on the average not great enough fully to compensate for the increase in cost. The main justification for the *status quo* with respect to such costs will therefore often take the line that the free-enterprise or capitalist system is on the whole the economic system we want, and that the system won't operate at all, or continue to be "free enterprise," unless it continues its traditional policies of advertising and sales promotion. This is the belief advanced by certain representatives to the federal Congress, when they maintain that any opposition to advertising in its present form is "un-American" or even "communistic" and aims at destroying the enterprise system. The issue thus raised is left open for the student's consideratior

PRICING RESULTS IN DIFFERENTIATED OLIGOPOLY—SUMMARY

What can be said of the pricing results in all industries with few sellers and differentiated products? Generalizations must rest largely on direct observation, since there is a considerable range of logical possibilities. Recognizing this, we suggest the following summary of the preceding arguments.

First, in most such oligopolies there is a tendency toward some excess of price over marginal cost. Output is not extended until marginal cost equals price. This is because there is a tendency toward monopolistic restriction of output both by oligopolistic industries and by individual oligopolists.[10] However, the degree of restriction relative to a given industry demand curve possibly tends to be less than in single-firm monopoly, because of the pressure of entry and of the reaction of established sellers to this pressure by lowering price and extending output. In fact, there may easily be extreme cases where this reaction leads to marginal cost in excess of price.

Second, the relation of price to average cost and the size of profits are generally uncertain, although more specific predictions may be made if we separate various oligopolies according to the ease of entry of new sellers, the vigor of the antitrust-law enforcement to which they are subject, and other aspects of market environment. Observation suggests that in a significant group of oligopolies supernormal or excess profits are ordinarily earned, but that in another group profits are low. In almost all differentiated oligopolies, severe price reductions and very low profits or losses may characterize depressions. The share of income going to profits under oligopoly probably exceeds that which would be so distributed under pure competition but is ordinarily smaller than it would be if single-firm monopolies occupied the same industries.

Third, selling costs per unit of product are probably larger in differentiated oligopoly than in any other market category. Such costs tend to be reflected in prices. This sort of market

[10] And, as in the case of monopoly, the attained aggregate cost of a given output may exceed the minimum attainable aggregate cost of that output, though not by any uniquely predictable amount. Cf. page 193.

organization favors the diversion of a significant proportion of productive resources to employment in product variation, advertising, and sales promotion. In return, the economy gets rapid product development, frequent style changes, a great variety of products, elaborate distributive service, and a means of subsidizing the press and radio. The cost of all these benefits, because of waste motion, tends to be relatively high.

Fourth, oligopoly prices ordinarily tend to be rigid over time. As indicated in the previous chapter, the effect of this phenomenon on economic stability is uncertain. For limited periods, oligopoly prices may be very unstable and flexible, especially during business depressions.

Fifth, there is observable no general tendency regarding size or rate of utilization of plant. The pressure of rivalry makes it somewhat more likely that plants will be of optimum size than in single-firm monopoly, but the competition is not of such a character as to make optimum size at all inevitable.

The relevant questions, of course, are, first, to what extent attained aggregate costs of actual industry output exceed the attainable minimum, and, second (even if the aggregate cost of attained output is minimized), to what extent average costs are raised above the lowest levels consistent with a desirable volume of industry output. Some departures from the minimum attainable aggregate costs of actual output occur in oligopoly, because of excessive or of insufficient entry, but these departures tend to be variable from industry to industry. Unit costs of production may be further raised in socially undesirable fashion by restriction of output below socially desirable levels—in cases of this sort where firms could reduce unit costs by extending output —but again the occurrence and the extent of the discrepancy are not uniquely predictable. In sum, there are departures from desirable efficiency, but not of an entirely systematic sort.

Progressiveness with differentiated oligopoly is not susceptible of simple characterization. On one hand, active nonprice rivalry, which seems fairly common in this sort of market, is conducive to very rapid progress in the adoption of new products and the improvement of existing products. Progress of this sort is certainly more rapid than in single-firm monopoly and probably more rapid than it would be under any other sort of market

organization. This seems to be the case basically because a degree of collusive monopolistic price control is coupled with fairly active rivalry on the nonprice level. The pressure of price rivalry is not strong enough to place a severe limit on selling expenses. At the same time, collusive oligopolists have the same general tendency as the single-firm monopolist to go slow in the adoption of new techniques. The net result will probably depend on the extensiveness of collusion and on the effectiveness of new entry in forcing the adoption of new developments.

There is a good deal about oligopoly pricing, however, which can be revealed only by detailed investigation of a large number of industries. This sort of market structure offers a wide scope to deliberative price policy by business management. The sorts of policies adopted, especially in response to cyclical and secular movements in income, require much study. Lacking the results of such study, we can only speculate about further behavior patterns.

PURE OLIGOPOLY

Our discussion of oligopolistic industries has thus far emphasized instances where the rival sellers have products which are significantly differentiated. This category is very important, including many of the industries producing consumer's goods. There is also a large group of industries with few sellers, however, in which product differentiation is insignificant or slight. These are mainly industries making producer's goods, which industrial buyers purchase, usually to specification and without much attention to brand or maker. The products of rival sellers in such industries are approximately standardized or homogeneous—they are perfect or almost perfect substitutes. The steel, copper, lead, cement, industrial alcohol, rayon, explosives, and processed building materials industries are some of those in this *pure oligopoly* category. Price policy and price results in this category are generically similar to those in differentiated oligopoly. But there are sufficient differences to merit some separate attention.

The setting for price-output determination in pure oligopoly resembles that for differentiated oligopoly in that each of the several sellers in a market for some product has a significant

share of the market, so that their pricing decisions become directly interdependent. In the steel market, for example, the largest firm (U. S. Steel) controls about 40 percent of the producing capacity of the industry, the largest five firms control about 75 percent, and the largest ten about 90 percent. If any of the first five firms were to announce a price cut of $5 per ton below the going level on all principal products of the industry, buyers would turn to it at once and place orders to the limit of its capacity. This would reduce the orders of at least some rival sellers by a large enough amount to cause them to retaliate to regain the volume, and price cuts would soon become general. None of the larger sellers, therefore, can make an *independent* price cut without inducing a chain of retaliation, nor can any such firm make an independent price increase and hope to maintain any sales volume unless it can induce its rivals to increase their prices also. Any one smaller seller would be unable to sell above the price of his major rivals. He might be able to make a price cut without engendering retaliation, because he could not supply enough at the lower price to reduce the sales of any rival greatly. But concurrent cuts by several of the smaller sellers would induce retaliation from major firms, so the smaller sellers as a group are in the position of any larger seller with respect to independent price changes. On the postulate of strictly independent action and no developed convention of concurrence on price changes, no large seller can know for certain how his rivals will react to his price cuts or increases, and thus on this assumption no such seller has a determinate demand schedule for his own output. Price changes undertaken independently will have unpredictable results on sales. Any smaller seller's demand at prices below the general industry level might be quasi-determinate, but as a group the smaller sellers have a demand which is likewise indeterminate. In these ways, the steel market (approximating pure oligopoly) resembles the automobile market (differentiated oligopoly).

The lack of significant product differentiation in pure oligopoly endows this market category with distinct attributes, first because it makes the interdependence of rival product prices much closer. In the substantial absence of differentiation, all rival prices must be identical (subject to the minor exceptions

noted), and no generally announced price differentials can be sustained. Thus if the going price for cold-rolled sheet steel were $70 per ton, no seller of steel could maintain his market for this product if he charged perceptibly more. If any important seller cut perceptibly below $70, he would take enough business away from others to bring them with him. This is because cold-rolled sheet is made to specification, and a buyer would as soon take the product from one seller as another. In pure oligopoly, therefore, the rival sellers must be able to keep their several prices close to identical if the price is to be stable and unaffected by retaliatory price adjustments. Where freight is an important portion of price, sellers must be able to maintain identical delivered prices at principal destinations if market stability is to be maintained. In pure oligopoly generally, there is little or no "slack" for individual price variations.

A concomitant of this is that in pure oligopoly the seller has relatively little opportunity to employ product variation and selling costs to expand or protect his share of the market. If the products of rivals were perfectly homogeneous, of course, there could by definition be no product variation, and there would be no point to advertising or sales promotion. In practice, there is a slight differentiation and a very limited opportunity for sales promotion. In the steel market, for example, sales promotion through entertaining prospective large buyers and otherwise building "good will" may play a part. But changes in the "style" of cold-rolled sheet, or advertising in trade journals that Old Glory Steel is naturally the finer, tougher, more elastic type of steel would avail little. This is because the buyers from purely oligopolistic industries are industrial buyers: they buy to measure and to specification, and are relatively little affected by emotional appeals. In effect, because the buyers are price conscious and also buy to specification, product differentiation and sales promotion cannot become important. The results of this are that the scope of rivalry among sellers in pure oligopoly is narrowed and centered on price, and that effective agreements to eliminate price cutting therefore become much more important than in differentiated oligopoly.

Another result of the relative absence of product differentiation, together with the restricted importance of sales promotion,

is that the shares of the market secured by the various rival sellers are potentially quite unstable. In the extreme of pure oligopoly, with absolutely no product differentiation, the shares of the several rivals would be quite indeterminate even though their prices were identical. Anyone could sell the whole market if he built enough capacity and if cost conditions made it attractive for him to do so. The shares actually obtained would therefore be determined by price warfare, by inertia, by chance, or, more probably, by agreement. In practice, with slight product differentiation and some sales promotion, the rival sellers have some means of controlling given shares of the market, but these shares are still potentially very unstable and subject to uncontrolled fluctuation if price rivalry among the sellers emerges.

The determination of price and output in pure oligopoly is thus a simpler and more clearly focused problem than it is in differentiated oligopoly. Several sellers have products which are either perfect or very close substitutes, so that in effect they share a market demand for a single product. There are really no separate demands for their several products. The problem is one of market price, total output, and sharing of this output, relatively uncomplicated by nonprice rivalry of various sorts.

The outcome of pricing in this situation is logically indeterminate. Three main possibilities, however, are suggested by the formal theory of pure oligopoly.

1. The oligopolists pursue independent pricing policies, foolishly overlooking the reactions of their rivals and independently varying their individual prices to enhance their several positions. The result of this, it can be demonstrated, would be wildly fluctuating prices, unstable markets for all sellers, and possibly eventual elimination via price warring of all but the strongest seller. Experience and logic suggest that this ordinarily does not happen, as it offers unattractive profit prospects to all or most sellers in the market. It has occurred principally in two situations. During the inception of the merger movement in various industries, severe price warfare emerged or was deliberately employed to eliminate small rivals. Here it was followed by a more concentrated oligopoly market structure and more stable price policies. Antitrust-law enforcement has tended to

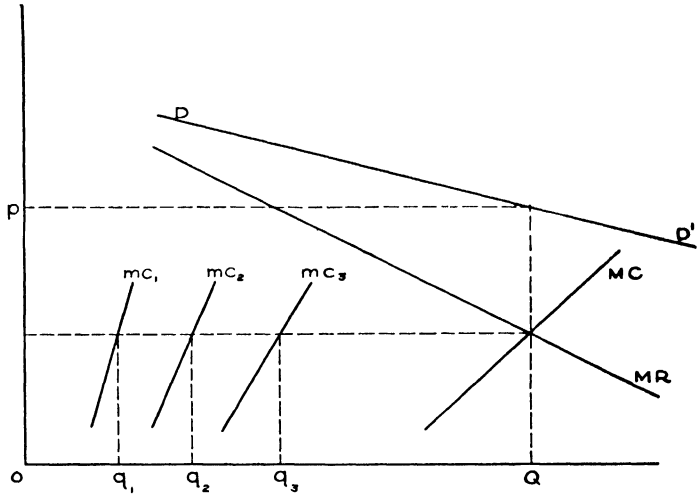

Figure 33

discourage predatory price cutting in later years, and to encourage "live-and-let-live" pricing. Independent price cutting also occurs periodically in severe depressions when price conventions or agreements temporarily break down. But it is not a normal way of pricing for oligopolistic industries.[11]

2. The oligopolists fully recognize their interdependence and make a full and concise agreement or cartel on price, combined output, and shares of this output. The possibility of such an agreement is fairly clear. Suppose, as in Figure 33, there are three (and only three) rival oligopolists with the respective marginal costs mc_1, mc_2, and mc_3, producing an undifferentiated refined mineral ingot. The aggregate of their marginal cost curves is MC, and shows the marginal cost of their joint production of successive industry outputs, each such output being produced at the lowest aggregate cost these sellers can attain. The market demand for the mineral ingot is DD', and the marginal receipts therefrom are MR. The oligopolists therefore agree to maximize their aggregate profit by producing the output OQ and selling at the price OP. They share the market in such wise that

[11] Cf. Chamberlin, *op. cit.*, Chap. 3, for a discussion of formal theories of oligopoly.

their several marginal costs are equal—that is, they produce respectively oq_1, oq_2, oq_3 (which necessarily total OQ).[12]

Such a "monopoly" price could be clearly limited by the threat of entry or of government interference, as in single-firm monopoly, and could be set for long-run advantage rather than the immediate maximization of short-run profits.

This model perhaps indicates a tendency inherent in oligopoly pricing, but such an exact determination of price, output, and market shares seems improbable. Even given the possibility of perfect collusion, oligopolists would probably not agree to share a market on the basis of equal marginal costs. If they do not, the shares are indeterminate, subject to the "power politics" of negotiation. Also, the best market price is then no longer the same for all sellers, and the price chosen will fall somewhere within a range of disagreement, depending again on bargaining among rivals. Further, such collusive agreements as that suggested are clearly contrary to antitrust law in the United States. Agreements made must therefore be *sub rosa* in character, legally unenforceable, and subject to defections and breakdowns. A legal expedient like price leadership will therefore probably be resorted to. This convention, which will also be subject to individual defections, will allow of less precise exploitation of the market demand, and will probably not fix market shares.

Still, oligopolies may tend toward the general sort of pricing indicated, subject to the imprecision introduced when price leadership must be relied upon.

3. The oligopolists may simply arrive at a fairly satisfactory price and all adhere to it for fear of "upsetting the applecart." In this case we get a very rigid price, which is not necessarily a monopoly price but may be lower or even higher than such a price. Various writers have probably overrated the practical importance of this pattern. There may be a few cases of this sort, but concurrent action via price leadership or secret collusion seems much more probable in pure oligopoly.

The central probability, therefore, as supported by logic and

[12] This solution is modified somewhat if one or more firms has falling marginal costs. Cf. Don Patinkin, "Multiple-Plant Firms, Cartels, and Imperfect Competition," *Quarterly Journal of Economics*, February 1947; and W. Leontief, "Comment," *ibid.*, August 1947.

by observation, is that of some approximation to monopoly pricing, but with market sharing subject to the lottery of power negotiation. Where price leadership is employed and all sellers post prices identical to that of the leader, market sharing may be further influenced by the independent granting of discounts and concessions to individual buyers by the various sellers. Such a pattern is clearly present, for example, in the steel industry. Pure oligopoly prices which are controlled by price leadership are also likely to be somewhat less flexible than a corresponding single-firm monopoly price.

The position in which the individual oligopolist arrives as the result of such concurrent pricing will depend upon further market considerations, especially (1) the ease of new entry to the industry and the manner in which established sellers have anticipated it, (2) the "trend of demand" in the industry, and (3) the degree of concentration in the oligopoly.

The closest approximation to profitable monopoly pricing is likely to be obtained where (1) new entry is effectively blocked, as by patent control or resource monopolization, (2) the demand for the industry is stable or expanding, and (3) there is a high degree of concentration. In this case a concord on price is easiest to maintain. The oligopolists can effectively hold down capacity, output, and costs while exploiting the market demand; and there is no shrinkage of demand to create excess capacity and higher average costs. Here we might expect that some collusive equation of marginal costs and marginal receipts would be made, and also that all or most of the oligopolists might make an excess profit. Thus the individual seller might be in some such position as in Figure 34, with fairly economical size and utilization, and price above both average and marginal costs.

Where entry has been relatively easy, the result will depend upon how established sellers have anticipated its effects. If they have tried to maintain attractive excess profits, as above, new entry may have reduced their shares of the market until the typical seller's position is like that in Figure 35, with price equal to a high average cost but above marginal cost. This situation may also result if an oligopoly industry has been faced with a secularly shrinking demand.

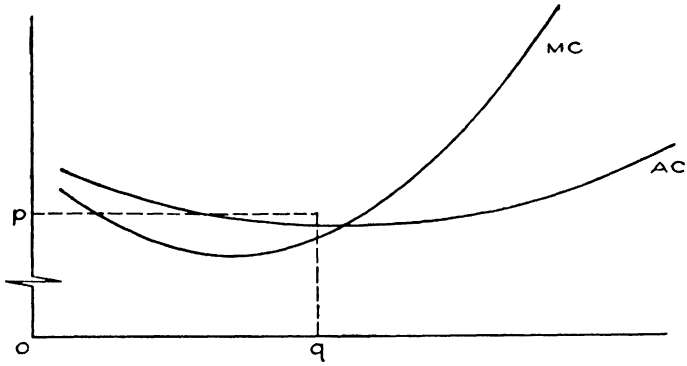

Figure 34

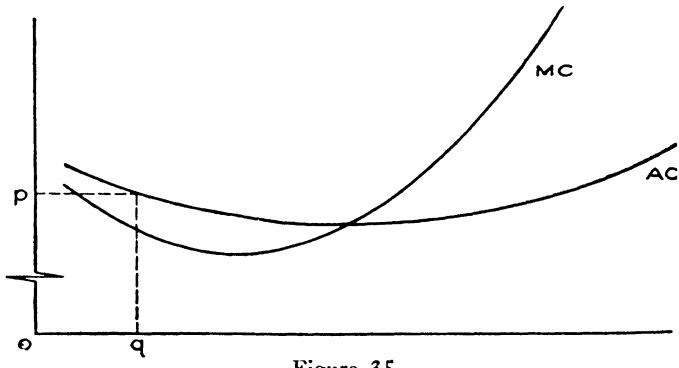

Figure 35

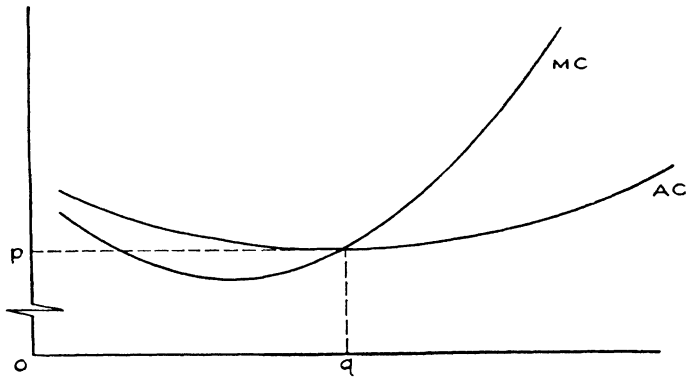

Figure 36

If new entry has been anticipated and forestalled by a low-price policy, the typical position may *approach* that in Figure 36, in which case some approximation to purely competitive results *could* occur.

Finally, in oligopolies with a low degree of concentration—for example, if there are twenty or thirty sellers of importance—the tendency for high-price concords to be disrupted by price cutting is strong. Price may in any event lie at lower levels. This result is by no means inevitable, however.

The price results emergent from pure oligopoly, though in strict logic indeterminate, may thus be characterized from observation and from a survey of probable tendencies as follows.

1. Price will tend to exceed average cost, allowing excess profits, in some of these industries, especially where concentration is high and entry is effectively blockaded. In others, because of entry, of anticipation of entry, or of declining demand, it may only cover average costs, allowing no more than sufficient returns to cover interest on investment. Short-run losses are quite possible in this case.

2. In some cases, usually where entry is blocked by institutional factors or forestalled by low-price policy, the level of production costs may be near the optimum.[13] Firms may be able to build to approximately optimum size and use plants at efficient rates. In other cases, where excessive entry has been attracted by high prices, or where demand is declining, oligopolists may collectively restrict their outputs so that costs are high and the scale and rate of use of plant are substantially short of the optimum, and most of the discrepancy may be socially undesirable.

3. Price will generally exceed marginal costs of production, because of monopolistic restriction of output. This excess may be quite small or even absent if the threat of entry leads to low-price policy, however.[14]

[13] Although with few firms the possibility of attainment of precise optima coincident with minimizing the aggregate cost of attained industry output would be only coincidental, successive increase in the appropriate number of firms makes closer approximation possible.

[14] And attained aggregate cost of a given output may be variously related to the minimum attainable aggregate cost.

4. Selling costs and costs of product variation will be low or absent. Hence total costs to be covered by price will generally be lower than in differentiated oligopoly.

5. Price rigidity is a common phenomenon.

6. Occasional periods of unstable or fluctuating price will occur in many cases, especially in severe business depressions.

The impact of such pricing on the economy should be fairly clear from our previous discussions of this matter. So far as prices exceed average costs and permit excess profits, a distortion of income distribution similar to that ascribed to monopoly occurs. The advantages and disadvantages of such profits are similar to those attributed to monopoly profits. When a chronic excess of firms develops, it raises costs and may raise prices. It is significant that, by monopolistic restriction of output, sellers in oligopoly can continue to earn a normal profit on excessive investment, when in an industry in pure competition this would not be possible.

The excess of price over marginal cost, common in differing degree to most oligopolies, has a significance of the same sort as discussed in connection with single-firm monopoly. Selling costs in pure oligopoly tend to be low, and most product variation is absent. Average total costs thus tend to be lower in this sort of industry than in differentiated oligopoly, and rivalry does not result in the diversion of so many resources to selling as compared with producing the product. Oligopolistic price rigidity may have some significance for business cycle behavior. The tendency to progressiveness in technique and product in pure oligopoly is probably smaller than in differentiated oligopoly, since nonprice rivalry is less important and collusion tends to be more comprehensive. Such progress may be more rapid in pure oligopoly than in single-firm monopoly, however, to the extent that new entry is instrumental in hastening adoption of new techniques.

OLIGOPOLY AND THE WORKING OF THE ECONOMY

Most of our industrial markets are oligopolies, either of the "differentiated" or "pure" variety. The former are found mostly in consumer's goods and the latter largely in producer's goods.

It is therefore well to pause and consider what this sort of market organization implies for the performance of the price system in ordering economic activity. With this, as with any sort of free-enterprise system, the pursuit of profit by various sellers will succeed on the one hand in allocating production among various lines *roughly* according to buyers' demands, and on the other in distributing income in some way. (There are further matters, however, to be discussed concerning income distribution.) But in an oligopolistic system, these functions will not be performed in any clear, simple, and definite way, nor necessarily in an ideal way. Their performance is subject to a considerable range of uncertainty, which arises out of the relatively indeterminate behavior of individual firms in oligopoly conditions. Within this range, the deliberated price policies of individual businessmen may play a large part in determining the behavior of price, output, and product.

Production and price in oligopoly are only distantly governed by the "invisible hand of the market." They are directly governed, and within a considerable free range determined, by the executives of principal business firms. Depending partly on strategic aspects of market structure and partly on the judgment of the responsible businessmen, price may be higher or lower and output smaller or larger within a considerable range.

Abstract theory probably cannot tell us exactly what sorts of price policies will emerge in oligopoly. To learn this, a detailed study of such price policies and of associated behavior is necessary. Theory plus general observation can suggest, however, certain tentative generalities about the price system under oligopoly. One is that there is a persistent, though moderated, tendency toward monopolistic restriction of output operating in varying degrees in most oligopolistic industries, facilitated by various collusive arrangements or conventions of price-making designed to secure concurrent action by rival firms. This results in some cases in excess profits, and in some others in the attraction of excess capacity. In yet others, low-price policies are followed and excess profits do not necessarily result. Whatever the result, it is not automatic.

Oligopolistic restriction also results in varying degrees of discrepancy between price and marginal cost, with a rather confus-

ing impact on the allocation of resources among uses. It may also make possible the preservation of normal earnings on excessive investment, or of earnings over and above the costs of wasteful sales promotion and product variation. The price system may be prevented from weeding out inefficiency as it would under more vigorous price competition. Another tendency in oligopolies is toward relatively rigid pricing policies, partly as the result of the instability which frequent price changes may generate. When many prices are quite rigid over the business cycle, the course of the cycle may be affected perceptibly.

Maintenance of rigid prices may be associated with a policy of setting prices sufficiently high to cover average costs at low and inefficient rates of output and therefore to yield excess profits in good times. This will cause a systematic cyclical fluctuation in the distribution of income.

Another and very important tendency in a large sector of oligopolies—where products are differentiated—is toward large selling and product-variation costs. In effect, a fair proportion of productive resources are thus devoted to promoting the sales of products, to varying and developing them, and to providing variety. Under oligopoly the proportion of resources thus used may be excessive from the standpoint of consumer welfare. The proportion which is used is not automatically limited by any market-price mechanism, but is ruled by the judgment of sellers in devising their price policies.

The effect on over-all output and employment of having the majority of industries organized as oligopolies is of the same general order as that attributed to a world of monopolies. There is some tendency to output restriction with a given ratio of money purchasing power to factor prices, but an adjustment of factor prices downward relative to purchasing power *may* allow the impact of the restrictions to be absorbed in lower factor prices and higher profits rather than in unemployment. In any event, the degree of restrictiveness is, because of easier entry, probably somewhat less than that attributed to single-firm monopoly.

In oligopoly the size of output, the relation of price to cost, the sorts of products and their quality, the proportion of re-

sources devoted to selling, and the proportions in which various products are produced are within a wide range determined by the deliberated decisions of large business organizations, and by the character of the pricing agreements or conventions they adopt or follow. The resulting pattern of economic activity may vary considerably and not necessarily in a desirable direction from any definable ideal pattern.

SUPPLEMENTARY READINGS

EDWARD H. CHAMBERLIN, *The Theory of Monopolistic Competition*, Chap. 3.

H. S. DENNISON AND J. K. GALBRAITH, *Modern Competition and Business Policy*, New York: Oxford University Press, 1938.

WALTON HAMILTON, *Price and Price Policies*, New York. McGraw-Hill Book Company, 1938.

E. G. NOURSE AND H. B. DRURY, *Industrial Price Policies and Economic Progress*, Washington, The Brookings Institution, 1938.

CHAPTER 7

THE EFFECTS OF CONCENTRATED BUYING

To this point all of our discussions of price determination have rested on one important implicit assumption—that wherever goods are sold, there are *many buyers,* none of whom buys a significant proportion of the total output of an industry. In effect, we have assumed that whatever the degree of concentration among sellers, the structure of the buying market is atomistic. For the great bulk of markets this is a fair and reasonable assumption. Practically all consumer's goods are sold to a multitude of small buyers, since in the nature of things consumers are many and their individual purchases are small. A considerable part of producer's goods are also sold to many buyers, none of whom buys enough to influence the market price seriously, and for those also our assumption is correct. But for certain producer's goods the buying market is concentrated or, in other words, dominated by a few large buyers. Observation and common sense suggest that when this market situation is found, the determination of price may follow a course somewhat different than that heretofore described. We must therefore pause to consider the peculiarities of pricing in industries with concentrated buying markets.

Where a good is sold to very many small buyers, no one of them buys enough that he can hope to influence the market price of the good. Each buyer necessarily accepts the going price and

buys whatever his financial means and his preferences dictate, since he cannot perceptibly influence a going price even by refusing to buy altogether. The student will readily recognize himself to be in just this position in buying any consumer's good. Such a small buyer is in a position with respect to price similar to that of the individual seller in pure competition. When all of the many buyers in a market must take this attitude and simply adjust their purchases in accordance with the going price for a good, the conventional market demand curve for the good emerges. It shows buyers as a group unqualifiedly ready to take certain amounts at certain prices, without regard to the conditions of supply. Assuming heretofore that there are always many small buyers, we have properly proceeded on the supposition that sellers are faced with certain demand curves for their products. These curves show the essentially competitive offers of buyers at various prices—offers which are made without regard to the costs of supplying the goods. This has given us a perfectly proper analysis for practically all consumer's-goods industries, and for a majority of producer's-goods industries.

Similarly, we have constructed the cost curves—*i.e.*, the relation of cost to output—for any seller on the supposition that he buys labor, equipment, materials, etc., also under competitive conditions, where he is such a small buyer that he cannot perceptibly influence the prices of what he buys by varying his purchases. Thus we have assumed that costs are calculated by sellers who in buying take the prices of productive factors as given. This is quite proper for the majority of cases.

When there are only a few buyers, however, the supposition that they will simply make certain purchases at certain prices, without paying any attention to the cost of supply or without attempting to bargain for a low price, is no longer necessarily valid. A large buyer, instead of accepting the market price of something he buys as outside his control, will ordinarily negotiate for a price or attempt otherwise to influence the price at which he buys. He can do this because he buys enough that by restricting his purchases, or threatening to do so, he can affect the welfare of sellers and therefore force them to bargain.

The general effect of fewness of buyers on price is evident to anyone. It will tend to result in lower buying prices. But it

should be possible to say more than this about concentrated buying. In fact, two significant results appear. First, the sellers supplying a market of a few buyers will not have a given market demand for their output but will rather be faced with specific bargaining offers which buyers make for their outputs. Second, firms buying labor and materials with which to produce will recognize that the prices of these things will vary as they buy more and extend their outputs, and will take account of this influence on the relation of costs to outputs. We shall therefore analyze first the effect of concentrated buying on the price and output of the good bought, and second its effect on the cost curves of the buyers.

PRICING UNDER SIMPLE MONOPSONY—ONE BUYER SUPPLIED BY MANY SELLERS

Some insight into the character of concentrated buying may be gained by considering the extreme case of a single buyer of the entire output of a good, or *monopsony*. Suppose that in a given colony of an imperial nation, a trading company is given a government monopoly on all commerce, and therefore is the sole possible buyer for the tobacco crop of the colony, which is produced by several hundred small independent farmers. How will it regulate its purchases? In general it will try to act in such a way as to maximize its profit from buying and selling tobacco, taking account not only of what the tobacco is worth to it for resale, *but also of the varying cost of supply of the tobacco from the farmers*. The character of its calculation may be made clear by Figure 37. Here *dd'* represents its tentative (but not effective) "demand curve" for tobacco, which shows the amounts of tobacco it would purchase at alternative prices *if it had no control over price*. Since it is a trading company, this demand curve is presumably derived from, and essentially a reflection of, the demand for tobacco by the people to whom it sells tobacco. (The character of this derivation will be discussed at a later point.) Each of the prices shown on *dd'* represents the net value to the monopsonist of an increment to his purchases of tobacco at the corresponding quantity point. Against this he will

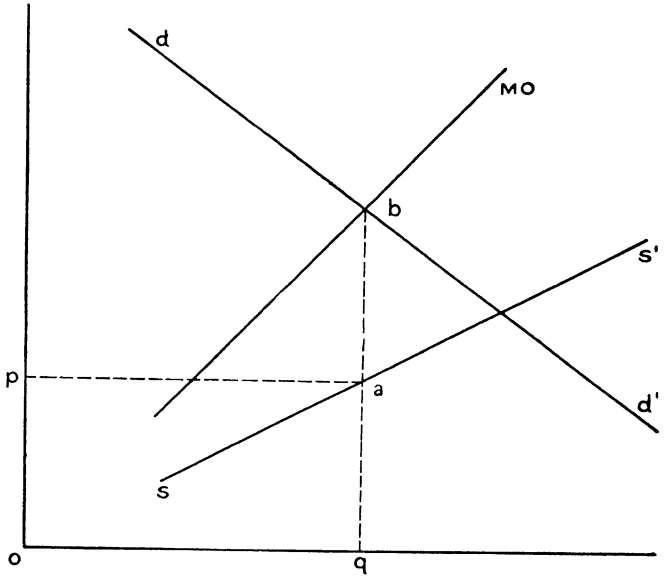

Figure 37

place the supply curve for tobacco *SS'*, which for a short period would be the summed marginal-cost curves of the many tobacco farmers. This is the conventional short-run supply curve for an industry of sellers in pure competition, and shows unequivocally the amounts sellers stand ready to supply at various prices.

Now if the demand curve were that of many small buyers, or alternatively if the monopsonist took no account of the varying cost of supply, price and quantity for raw tobacco from the farm would be determined at the intersection of *SS'* and *dd'*. But the monopsonist, as sole buyer, will naturally take account of the minimum prices, *shown on SS'*, at which various quantities of tobacco can be bought, and also of how these minimum prices increase as he extends his purchases of tobacco. If the supply curve slopes upward to the right, so that supply prices increase with output, the monopsonist will find that by increasing his purchases he raises the total necessary outlay for tobacco very steeply. Thus if the supply curve shows the price-quantity relation indicated in columns 1 and 2 below, the additional or *marginal outlay* on tobacco by the buyer rises as shown in column 4.

(1) Price, in guineas	(2) Tons supplied	(3) Total outlay, in guineas	(4) Marginal outlay of the buyer, in guineas
10	5	50	—
11	6	66	16
12	7	84	18
13	8	104	20

Thus the tobacco monopsonist will say to himself, "I can buy 5 tons of tobacco at 10 guineas, or 6 tons at 11 guineas, *in which case the sixth ton costs me 16 guineas more, since in getting one more ton, I raise the price of all tobacco.* Hence I will not consider hereafter simply the supply price of various amounts of tobacco, but rather the *marginal outlays* incurred in adding to my purchases of tobacco." In this case the monopsonist evidently will refer to a *marginal-outlay curve,* drawn to the supply curve, which shows the rate of increase of his total outlay with increase in his purchases (always assuming he buys any amount at its minimum supply price). This is the curve *MO* in Figure 37 (a prototype of column 4 in the preceding table).

To maximize his return on tobacco operations, the monopsonist will evidently purchase the quantity *oq*, for which his marginal outlay just equals the net value of the added tobacco. For this amount he will pay the minimum price *op*. He will earn an "exploitative margin" of *ab* per unit. This is essentially his monopsony earning from full control of all buying.[1]

Where the supply price to a monopsonist increases with increasing supply, as shown above and as common to many industries, the effect of monopsony is obviously both to lower the buying price and to reduce the output from the levels associated with competitive buying. Even if the monopsonist in turn sells in a competitive buying market, the reduction in price will not be passed on to the next buyers (for example, tobacco users) but

[1] See Joan Robinson, *The Economics of Imperfect Competition* (London, Macmillan and Co., Ltd., 1933), Chap. 18.

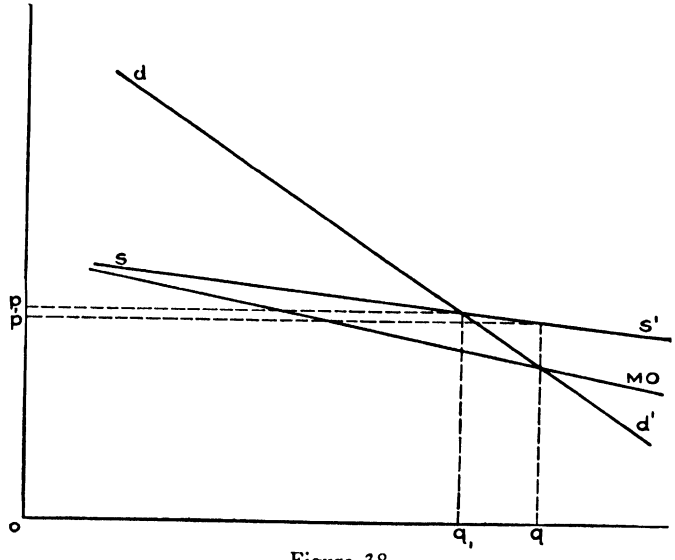

Figure 38

will be retained as an excessive earning of the monopsonist. The price to the ultimate users of the good tend to be higher and the volume smaller than under competitive conditions, since the supply of the basic raw material has been curtailed. Only the basic suppliers' price and income have been forced down, and this to the sole advantage of the monopsonist.

The result is not perfectly general, however, but applies only where the competitive supply curve for the good purchased by the monopsonist evidences increasing cost with increasing supply. Where the competitive supply curve is horizontal (constant costs) no change from the competitive buying price will result from monopsony, since the marginal outlay curve of the monopsonist coincides with the supply curve (*SS'* and *MO* are the same). Where the supply curve slopes downward to the right, evidencing declining supply price with increasing output, as may occur in some cases, the effect will be to extend the output beyond the competitive level as well as to lower the price. Thus if *dd'* and *SS'* are as shown in Figure 38, *MO* will be below *SS'* and the price and output will be at *op* and *oq*, whereas competitive price and output would be op_1 and oq_1. Monopsony thus

can result in a volume of production the same as or greater than that associated with competitive buying, but it will evidently never result in a higher than competitive buying price.

Where the monopsonist buys from a competitive industry, however, it is almost certain that the short-run supply curve will show increasing cost, and it is very probable that the long-run supply curve will also slope upward to the right. This is generally true of agricultural industries, from which concentrated buying is most common. Thus the probable tendency in monopsonistic buying from a competitive industry is for restricted output, increased price to consumers, lowered price to the basic suppliers, and excess buying profits to the monopsonist.

All this reasoning, however, proceeds on the assumption of a given demand by the monopsonist and a given up-sloping supply curve for the supplying industry. If these are given, then indeed monopsony results in lower price and output than competitive buying. To ascertain whether they may be safely regarded as given, and also what effect monopsony in one industry may have on the economy as a whole, we should carefully investigate the implications of a given up-sloping supply curve for the monopsonized supplying industry. This supply curve suggests that if the price for the output of the industry is cut below the competitive level, the industry will continue to produce but will produce necessarily less at successively lower prices. Such behavior in turn implies that at lower prices the industry will employ fewer resources, reducing its employment of labor, land, and capital, and that the resources released will therefore either become idle or will find employment in other industries. The effect of the monopsony on the economy as a whole will evidently depend upon how elastic the supply of the monopsonized industry is, and also on whether the resources released as industry output declines become idle or find alternative employment.

At one extreme, all the resources might be perfectly mobile, in the sense that they would move freely to other industries if their rates of pay in the monopsonized industry were cut at all below the prevailing competitive rates for the economy as a whole. But in this event, the industry would have a perfectly

elastic long-run supply curve (costs could not be driven down at all) and the monopsonist would have no monopsony power whatever (*MO* would be identical with *SS'*). The up-sloping supply curve evidently supposes that at least one resource employed by the supplying industry is immobile (cannot be employed elsewhere) or imperfectly mobile (will accept lower than the prevailing economy-wide price before leaving this industry), so that some or all of it will accept lower prices as the industry demand for it is reduced, and thus give the industry lower costs at smaller outputs. It also supposes that one or more factors used by the industry either is mobile, so that units of it will leave employment in the industry rather than accept successively lower prices, or will prefer idleness to employment at lower prices. (If all factors were entirely immobile and shunned idleness, the industry supply curve would be perfectly inelastic and the monopsonist would employ them all at a zero price.) The typical up-sloping supply curve which gives the monopsonist his advantage thus rests on the *imperfect or partial mobility* of factors away from employment in the monopsonized industry. It is accompanied either by potential mobility to other industries or mobility to idleness of at least some of the resources the industry uses.

From this it follows that the impact of a monopsony on the economy as a whole *may* be entirely upon allocation and income distribution; where the monopsonistic restriction takes advantage of imperfectly mobile resources to pay them lower prices, but where all disemployed resources move to other industries. Here, the industry output is restricted, incomes of imperfectly mobile factors are reduced, and monopsonistic profits increased, but the shortage of output in this industry is compensated by an increase in that of other industries. The principal effect on total output is on its composition—on the allocation of resources among uses. It is also possible that the impact of monopsony may be entirely on total output and income distribution. That is, the monopsonistic restriction idles resources which refuse lower rates of pay, but no resources move to other industries. Then the loss of output in the monopsonized industry is not counterbalanced, and the aggregate output of the economy is lessened. More probably, monopsony will have an admixture of

these two effects. By lowering price and output in the monopsonized industry and increasing its profits, it will drive some resources to other industries, where they increase other outputs, and will drive some to idleness. It will thus have effects both on aggregate employment and output and on the allocation of resources among uses. It may be worthwhile noting that very immobile resources under monopsony will mean very inelastic industry supplies, so that monopsony can drive price very low with little restriction of output.

It may also be pertinent to reflect upon the potentialities of a world of monopsonies—of having very many industries of an economy each subject to monopsony buying. Here the possibility of mobility of resources among industries is greatly reduced if not eliminated. The combination of many individual monopsony restrictions, therefore, does not necessarily mean an aggregate restriction equal to their sum. Under general monopsony pressure, the prices of all employed factors may be forced downward, so that all costs and supply curves lie at lower levels. If this adjustment can take place relative to a given aggregate money purchasing power, aggregate output will not be much restricted, and the entire burden of monopsony may fall upon income distribution (a shift from wages and rents to excess profits), except so far as employable resources prefer idleness to low rates of pay. This model may be appropriate to colonial economies subject to monopsonistic employment of the bulk of the labor force. It should be noted, however, that such a slight effect upon employment rests upon the possibility of money factor prices being forced down while purchasing power is sustained, and this may not be accepted as a foregone conclusion.

The discussion so far has employed the convenient device of a "tentative" demand curve for the monopsonist, representing the net value to him of various quantities of the good he purchases. This may now be replaced by using a more general analysis, in which we refer to the demand *for what the monopsonist sells* (and from which this tentative demand for what he buys is evidently derived) and to his costs of production. In this way also we may recognize explicitly the effect of monopsony on the cost curves of producing firms with monopsony power.

Let us refer again to our monopsonistic buyer of the tobacco crop of a colony. The tentative demand of any such buyer for a raw material is evidently not a consumer's demand, or demand for consumption by the buyer (since there will ordinarily be many consumers of a good) but rather a producer's or merchant's demand, based upon or derived from the demand which the merchant or producer will meet when he resells the product. The monopsonistic tobacco buyer intends to resell the tobacco, either before or after processing it or using it in production of cigarettes and cut plug; the amount he is willing to pay for it depends upon the demand curve for tobacco in his selling market. It follows that we can trace his policy as a buyer and a seller together by placing his costs, including the cost of leaf tobacco, against the demand for his offering as a seller.

There are numerous possible variant situations which could be illustrated—here he will choose a very simple model. Let us suppose that the monopsonistic buyer of tobacco is also a single-firm monopolistic seller of tobacco, with a given demand curve for his output—DD' in Figure 39. Let us also suppose that he is simply a trader or merchant, so that he just buys and sells tobacco, with no additional production costs. The supply curve for tobacco as he buys it is SS', the same as in Figure 38, and the marginal outlay drawn thereto is similarly MO. Now the supply price of tobacco to the monopsonist represents his entire cost of production, and it is therefore evident that SS', showing the buying price per unit for various amounts is the monopsonist's *average-cost curve* ($AC = SS'$), and that the marginal-outlay curve is his marginal-cost curve ($MC = MO$). In effect, his marginal cost of production rises more steeply than a competitive buyer's would because the unit price of materials rises as he buys more. The curve MR represents the marginal receipts corresponding to the demand curve (DD') in his selling market, and it is thus equivalent to his own derived demand for buying tobacco ($dd' = MR$) as shown originally in Figure 37. The monopsonist-monopolist thus sets his purchases and sales at the level where MO (MC) intersects dd' (MR). His volume is oq (as shown also in Figure 37), his buying price is OP (also shown in Figure 37), and his selling price is OP'. He earns an exploitative margin or excess profit

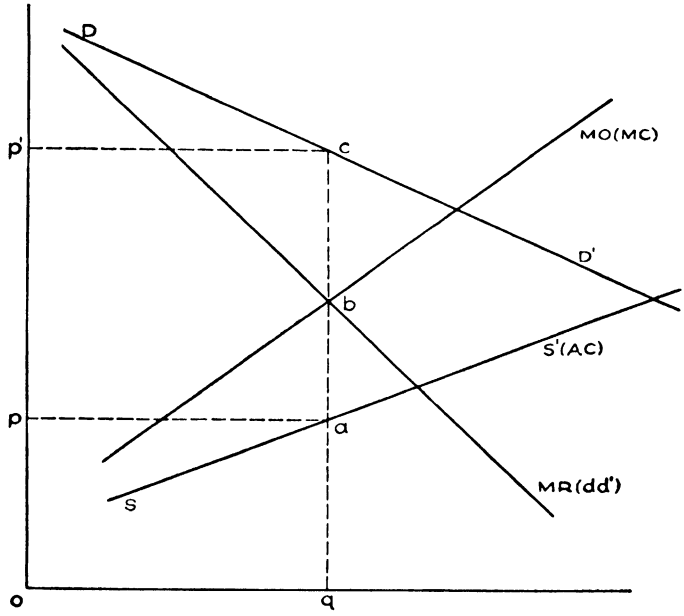

Figure 39

of *ac* per unit, made up of a monopsonistic profit *ab* (also shown in Figure 37), and a monopolistic profit of *bc*. Thus we see the derivation of the monopsonist's demand for what he buys (from the marginal receipts from the demand in his selling market), and his combined operation as a buyer and seller.

The effect of monopsony power on the cost curve of a producer possessing it may be significant in many instances. Where the prices of the things a producer buys or hires rise in response to an increase in his demand for them, both his average-cost and marginal-cost curves rise more steeply than they otherwise would, since they are influenced not only by varying proportions of factors but by the rising prices of these factors. Such monopsony thus tends virtually to restrict the outputs of producers possessing it, and hence to influence adversely the allocation of resources among alternative uses. We will discuss the effect of monopsony on allocation in a special note at the end of this chapter, designed for classes which wish to pursue the analysis of allocation a bit further.

A FEW BUYERS SUPPLIED BY MANY SELLERS—COLLUSIVE OLIGOPSONY

What is the significance of this analysis for pricing in actual business practice? Needless to say, there are virtually no actual cases of absolute monopsony in commodities; the preceding analysis cannot often be applied directly. Behavior similar to that just described may be found, however, where there are very *few* buyers of a competitively sold raw material, and where they exercise collusion to concur on a monopsony buying policy. Such concurrence may be obtained either by agreement or by buying-price leadership, whereby one of the buyers sets a low buying price which others follow. This is essentially the case of a collusive *oligopsony* (few buyers) buying in a competitive raw-material market. The principal opportunities for such buying are found where a few large processors buy a raw material from farmers or other small producers. Monopsony buying practices have been alleged to occur in the raw-tobacco market, where a few large cigarette makers buy most of the tobacco; in the central cattle market, dominated by a few large meat packers; and in the midcontinent crude-oil market, where a few large oil companies buy most of the crude petroleum. We cannot investigate the validity of these allegations here, but such behavior is theoretically quite feasible.

Where a collusive oligopsony operates effectively in buying competitively supplied raw materials, the effects will, of course, depend on the character of market supply. In the agricultural sphere at least, supply prices probably increase with output. Collusive oligopsony would then tend to restrict supply, lower the income of the farmer, and raise consumer prices. The raising of prices to final buyers, however, depends on further effective collusion among the processors in keeping up their selling prices and thus retaining their monopsony gains. It is quite possible that such firms would, perhaps because of the threat of new entry or for other reasons, pass on part or all of the advantage of low raw-material buying prices to consumers. The result here is quite uncertain and could best be ascertained by

detailed investigation. Where the oligopsony is successfully selfish, however, restriction of output, and an enhancement of the oligopsonists' share of income at the expense of both consumers and farmers, tend to result. It must, of course, be recognized that unrestricted rivalry in buying among an oligopsony of buyers (an improbable case) could have almost any price result.[2]

BILATERAL MONOPOLY AND BILATERAL OLIGOPOLY

An equally important case of concentrated buying is found where a few buyers acquire a good from an oligopoly of a few sellers. This situation occurs, for example, in the selling market of the rubber-tire industry, where a fairly concentrated oligopoly of tire manufacturers sells much of its output to the three major automobile companies and to a few mass distributors such as Sears, Roebuck, Montgomery Ward, and several auto-supply and service-station chains. It plays a significant role in the sale of important steel products, where an oligopoly of steel mills sells sheet and other steel to a few automobile firms, and tin plate to a few can manufacturers. There are several other important cases of this sort.

Where a few sellers supply a few buyers we have oligopoly plus oligopsony, or appropriately *bilateral oligopoly*. This evidently issues us into the environment of negotiation, bargaining, power maneuver, and economic warfare, where price may fall anywhere within rather wide limits.

The simple theoretical prototype of this confused situation is that of bilateral monopoly—one seller selling to one buyer. Precise analysis of this situation reveals the following. (1) The monopsonist wants a restricted monopsony output and a low price. (2) The monopolist wants a restricted (but different) output and a high monopoly price. (3) The result may fall at either limit if one party has dominant bargaining strength; or it may fall uncertainly between the limits; or the protagonists

[2] For an extended discussion of oligopsony, see William H. Nicholls, *Imperfect Competition within Agricultural Industries* (Ames, Iowa: Iowa State College Press, 1941), Chaps. 4-9.

may "get together," maximize their combined return, and divide the loot in some proportion.[3]

Much the same range of alternatives holds good for bilateral oligopoly, although the larger number of principals in the fight makes it less likely that it will be fixed. All sellers and buyers are unlikely to agree on a mutual price and division of the excess returns. In fact, the scales are fairly strongly tipped in favor of the buyers in such a struggle in bilateral oligopoly, since each of them can act independently to drive a hard bargain as he seeks supply, whereas to resist this pressure the sellers must have effective collusion on price and maintain it under duress. Thus large buyers from the rubber-tire industry during the 1930's continually kept their buying prices at very low levels, allowing subnormal profits to tire makers, through vigorous but apparently not collusive use of their bargaining power.

In general we must admit the several possibilities of collusion among buyers, collusion among sellers, or collusion of buyers and sellers together. If the sellers have effective collusion and present the buyers with a uniform fixed price regardless of volume of purchases (essentially a horizontal supply curve), the sellers may make a monopoly profit, whereas the buyers will be able to derive no advantage from bargaining and no "monopsony" distortions will occur. Effective collusion by buyers alone, on the other hand, may result in a sort of monopsony restriction on output and price, driving sellers' returns to a bare necessary minimum. Balanced power makes the result uncertain within a significant range.

The significance of bilateral oligopoly market situations in our economy, in sum, is largely to compound the uncertainty concerning the way in which resources will be allocated, goods priced, and incomes distributed. The student may already have observed that bilateral oligopoly or bilateral monopoly situations may frequently occur in the labor market. This will be discussed in a later chapter.

We have spoken in the two preceding chapters of the behavior of prices and outputs in an economy which is a world of monopo-

[3] See Nicholls, *op. cit.*, Chaps. 10 and 11.

lies or of oligopolies—where in most selling markets sellers possess a degree of monopoly power and attain some approximation to a monopolistic relation of marginal and average costs to price. The corresponding effects upon income distribution (in the direction of excess profits), upon aggregate output, and upon the allocation of resources among uses have been discussed. To appraise the behavior of the real economy, we must compound with the effects of monopoly and oligopoly in selling markets those of monopsony, oligopsony, and bilateral monopoly and bilateral oligopoly in buying markets. The additional effects of simple monopsony and oligopsony are fairly clear. They add to the distortion in the allocation of resources, further tend to create excess profits, and thus reduce other distributive shares. But they have an effect on aggregate output mainly in so far as there is immobility of resources from industry to industry or in so far as the effects upon income distribution affect employment. (Lower real prices of hired resources may result in fewer resources seeking employment, or the ratio of money purchasing power to prices may be adversely affected.)

Where there is bilateral monopoly or oligopoly in buying markets, however, the final outcome of the relation of prices to costs—and with it allocation, income distribution, and perhaps total employment—are made logically uncertain over a significant range, and an observation of actual behavior is the only reliable guide. With the markets for labor increasingly assuming bilateral-monopoly characteristics, and with a good deal of bilateral oligopoly in the markets for producer's goods, it is hazardous to draw logical deductions concerning price-output behavior in the economy unless we allow a very substantial margin for error.

MONOPSONY AND ALLOCATION—FURTHER REMARKS

We may here refer again, for the benefit of those interested in a further discussion of allocation, to our treatment of the effect of monopolistic pricing on the output of an industry and the allocation of resources among uses. In Chapter 5 (pp. 164–165) it was assumed that the monopolistic seller bought factors under competitive conditions, so that the prices he paid for fac-

tors did not vary in response to variations in his output. His marginal cost curve was thus drawn on the assumption of given factor prices, and it thus reflected at each point only the money value of the marginal real cost—of the real factors added to produce another unit of output. Even on this assumption, the monopolist restricts output, since he sets output where marginal cost so defined equals marginal receipts rather than extending output until such marginal cost equals price, as a competitive industry would. But now if the monopolistic seller is also a monopsonistic buyer, for whom the prices of purchased factors rise with his output, his marginal cost curve no longer is drawn on the assumption of given factor prices, and it no longer represents simply the money value of the marginal real cost. It represents this *plus the increment to money outlay on all* ("intramarginal") *factors employed prior to the instant increment in real cost*. The marginal cost curve of the monopolist-monopsonist thus lies above the curve which would show the money value of marginal real cost and rises more steeply. When the seller equates this marginal cost to marginal receipts, therefore, he chooses an even lower output and higher price than he would have if he had equated the money value of marginal real cost to marginal receipts. There is thus a *double* restriction of output—monopolistic restriction *and* monopsonistic restriction.

The cost curves shown in Figure 39 are not strictly comparable to those represented for the monopolist in Chapter 5, since the latter reflected a rising marginal real cost within the firm due to diminishing returns against a fixed management factor (and no rise in factor prices) whereas the former presuppose no such tendency within the firm (but rather constant marginal real costs to the firm—it always costs the firm one unit of tobacco bought to have one unit to sell). The only reason for the rise in money costs in Figure 39 is thus the rise in the price of tobacco.

Recognizing this, the following general comparison may nevertheless be drawn. The curve *AC* (*SS'*) in Figure 39 represents the money value of marginal real cost for the monopolist. (In this respect it is comparable with the *MC* curves drawn in Chapter 5, but not otherwise.) The *MO* (*MC*) curve represents the marginal cost of the monopolist including the effect of

induced factor-price increases for all units of factors used (the money value of marginal real costs plus the money cost of intramarginal factor-price increases). If the monopolist were to set output so that AC was equal to MR (which he would not) he would restrict output somewhat below the level where AC intersects DD' (the competitive level). He would set the money value of marginal real cost equal to marginal receipts. This much restriction was attributed to monopoly in Chapter 5, where the marginal cost curve was drawn on the assumption that factor prices were invariant, and thus reflected only the money value of marginal real cost. But the monopolist in Figure 39 in fact sets output where MC equals MR, since he is concerned mainly with the total rise in his money costs, and he thus restricts output below the level where AC equals MR and further raises price. The first output restriction [from (AC=DD') to (AC=MR)] may be referred to as *monopolistic* output restriction, and the second [from (AC=MR) to (MC=MR)] as *monopsonistic* restriction. Where the monopolist is also a monopsonist facing increasing factor prices with increasing output, the two restrictions are added and the total restriction increased. Since the criterion for competitive output and ideal allocation is that output be set so that AC intersects DD' and equals price (thus putting the money value of marginal real cost equal to price), it is evident that the departure from competitive output is greater when the monopolist has monopsony power of the sort mentioned than when he does not. The preceding does not exhaust all the variant cases of monopsony effects (*e.g.*, where the monopsonist's factor prices decline with increasing output) but may serve to indicate the general character of monopsony effects.

Applying the preceding to the allocation argument, it is clear that ideal allocation of resources as between a competitive industry and a monopsonist-monopolist industry, both facing rising factor prices with increasing output, would *not* be satisfied *even if the monopolist's marginal cost equaled price and the competitive industry's average cost* (money value of marginal real cost) *equaled price.* For since the monopolist's marginal cost is then greater than the money value of marginal real cost, the latter would still be less than price in the monopoly, whereas

it would equal price in the competitive industry. Assuming both industries draw on the same resources, and pay the same prices, it would still follow that the last unit of resources employed in the monopolistic industry yielded a product selling for more than that yielded by the last unit employed in the competitive industry. Ideal allocation requires that in each industry the yield of the last unit of resources be the same. Assuming common factor prices, this means that the money value of marginal real cost should equal price in each case, or at any rate be in the same proportion to price. Or, in effect, the marginal cost which is balanced against price in each case must be an industry marginal cost calculated to exclude any induced rise in the prices of intra-marginal units of factors. Monopsony may give rise to a restriction of output or distortion of allocation in addition to that charged to monopoly.

SUPPLEMENTARY READINGS

JOAN ROBINSON, *The Economics of Imperfect Competition*, London: Macmillan & Company, 1933, Chap. 18.

J. R. HICKS, "Annual Survey of Economic Theory: The Theory of Monopoly," *Econometrica*, vol. 3, pp. 215 ff.

WILLIAM H. NICHOLLS, *Imperfect Competition within Agricultural Industries*, Ames, Iowa: Iowa State College Press, 1941.

WILLIAM FELLNER, "Prices and Wages under Bilateral Monopoly," *Quarterly Journal of Economics*, August 1947.

CHAPTER 8

MARKETS IN MONOPOLISTIC COMPETITION

Our discussion of prices and price formation in selling markets has been concerned to this point with markets in pure competition, with single-firm monopolies, and with oligopolistic markets. The first two sorts of markets are relatively unimportant in the American economy, and a detailed investigation of them was justified mainly because they illustrate certain characteristics of the pricing system in an extreme and simple form. The theory of oligopoly pricing is actually very closely related to the real economy, since the great bulk of industrial markets have one sort or another of oligopolistic structure. One further type of market, however, remains to be discussed—the market in *monopolistic competition.* This designation is used narrowly here to refer to markets where there are many small sellers, and where their products are differentiated. In effect, it refers to competition within large groups of close- but not perfect-substitute products.

Monopolistic competition is not as important a market type as oligopoly in the American economy. Most of our mineral extractive industries, our basic processing industries, and our manufacturing and assembly industries have oligopolistic structures. But a significant number of situations of monopolistic competition are found in manufacturing industries, and the type is quite common in the distributive trades, especially in the retail-

ing of a considerable number of products. In the manufacturing field, monopolistic competition is found in the ladies'-dress industry, in shoes, in millinery, and in similar consumer's-goods industries where small-scale operations have proved to be economical. In the distributive field, the grocery stores, clothing stores, drug stores, electric-appliance stores, and the like in any locality constitute industries giving some approximation to monopolistic competition. The market type is therefore well worth some investigation.

An industry in monopolistic competition occurs when a "large" number of sellers produce close-substitute but not identical products. The number of sellers should be great enough, and the largest seller of the group small enough, that no one controls a significant proportion of the group market. In effect, each controls little enough that by extending or restricting his own sales he does not perceptibly affect the sales of any other individual seller. There is thus no *recognized* interdependence of the related sellers' prices or price policies. Any seller may raise or lower his price and reduce or extend his sales volume without eliciting a rivalrous reaction from any other seller in the group. The situation is thus like that of differentiated oligopoly, except that the number of sellers is large enough that there is no recognized interdependence of price policies, and that each seller pursues an independent course. The decisions of individual sellers will affect one another, but indirectly through a "market" reaction rather than directly through explicit rivalry of one sort or another.

This is the exact theoretical category of monopolistic competition. Not many of the industries which we have characterized as falling therein fulfill its conditions precisely. Although in certain industries there are many sellers of close-substitute but differentiated products, there is ordinarily some slight degree of recognized interdependence among them, or within certain subgroups of them. But this interdependence is unimportant enough that they *approximate* markets in monopolistic competition, and an examination of the strict theoretical type may reveal a good deal about them.

The exact theoretical case would occur, for example, if in a metropolitan area there were perhaps 100 small independent

grocery stores, selling similar lines of groceries, but with their products differentiated by service and location, where each was *equally* in competition with every one of his 99 rivals. Any price cuts which one grocer made would take business from his rivals, but in roughly equal proportions, so that any competitor, feeling only about 1/99th of the blow of the price cut, would suffer so little as not to notice it and therefore would not react. In the actual case, some of the 100 groceries would be large enough that they would feel the effect of each other's price changes and have some recognized interdependence. Also, closely neighboring stores, in the same shopping center or only a block or two apart, would have a degree of recognized interdependence. Monopolistic competition would thus be alloyed with a bit of oligopoly. But the recognized interdependence would be small enough that we may emphasize the other aspect of the matter.

PRICING IN MONOPOLISTIC COMPETITION

Price-output determination in monopolistic competition is relatively simple if we for the moment rule out both product variation and selling costs. We shall consider this simple case first. Here we have a large number of sellers producing close-substitute but differentiated products. The pattern of differentiation is frozen, so that the various sellers are not changing their products over time. Two additional considerations are implicit in the large number of small sellers: (1) entry to the market is evidently easy—entry, that is, of additional close substitutes, although not of identical duplicates of existing products; and (2) collusive action by so many sellers is practically impossible. Thus we might imagine an industry of 80 small manufacturers of cigarette lighters, differentiated by design and branding, but with all products for the moment assumed "static." No collusion is feasible, and it is very easy for additional firms to enter the field with different brands or designs.

In this event, each small seller has at any given time a demand curve for his product, showing the amounts of his product he can sell at each possible price, *provided his competitors' prices remain unchanged at their current levels.* This demand curve is

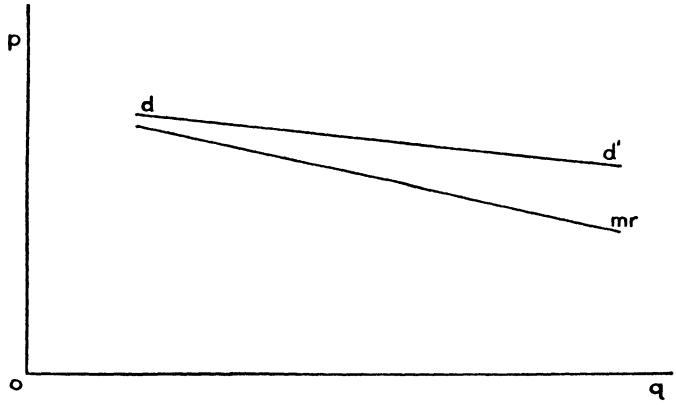

Figure 40

presumably sloping, but rather slightly. It is very elastic on both sides of the seller's current price, indicating that small reductions in his price will add greatly to his sales and small price increases will greatly restrict his sales, provided competitors' prices do not change. This great elasticity of demand stems from the fact that the various brands of cigarette lighters are rather close substitutes, so that a small price change for any one will have a big effect in its sales. The individual seller's demand curve, and the associated marginal-receipts curve, thus look like those in Figure 40, making him a sort of small monopolist.

We may properly inquire what meaning this curve has, since it rests on the tenuous assumption that all competitors' prices remain given while any one seller in question manipulates his price. The answer is that although the competitors' prices may change, they will not change *because* of price changes of the single seller. Therefore at any one time each seller, having no direct interdependence with others, can legitimately feel he has an independent demand curve for his own product, which he can exploit without inducing reactions which would cause it to shift about.

Each monopolistic competitor, therefore, acting independently (as in the assumed situation he will) makes his own adjustment of price and output to his current demand curve. That is, he maximizes his profit by choosing that output, *oq*, for which

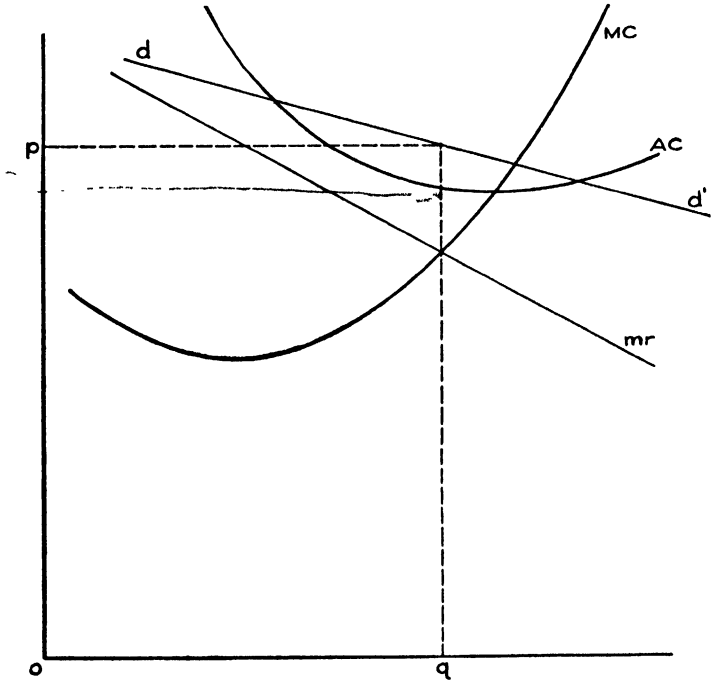

Figure 41

his marginal cost equals his marginal receipts, as in Figure 41. This may or may not give him an excess profit, although in the long run he would not operate at a loss.

We cannot continue to observe one seller at a time, however. Each one of the 80 sellers of cigarette lighters will continually strive to make precisely the same independent adjustment, selecting the best price-output combination for himself. And as all sellers do this together, *they will virtually codetermine the positions of their several demand curves.* In fact, these 80 demands are a family of demands, and the position of each of them depends on the prices charged by 79 other sellers. Their final positions therefore are not set until every seller has adjusted himself to his demand curve in successive positions until all sellers are maximizing profits along *mutually consistent* individual demand curves. In effect, a *group equilibrium* must be struck, *where the aggregate output of the 80 substitute products*

is equal to the aggregate demand at the existing family of prices. In such a group equilibrium, each seller is producing an output and charging a price which in terms of his own demand (dd') maximizes his profit, and this same action on the part of all sellers leaves the various individual demand curves in unchanged positions.

The welfare of individual sellers in this group equilibrium is not precisely determinate so long as we suppose the number of sellers to be arbitrarily given. In the short run, it could yield excess profits, normal profits, or losses, and in the long run, normal profits or excess profits could result. Thus the long-run equilibrium of all of the sellers, barring new entry, could be as shown in Figure 41. Here the typical seller is making a long-run excess profit per unit equal to the indicated gap between price and average cost. (We assume for the moment that the sellers are similar in their individual costs and in their individual demands.)

Should sellers generally make excess profits, however, this would, in the long run, attract new entry by additional firms. This entry, by dividing the market among more sellers, would in turn reduce the demand for the product of each, shifting his demand curve downward. Entry would presumably continue until each seller was in a position where his best price just covered average costs, as in Figure 42. (We assume for the moment that the relation of cost to demand is the same for all sellers.) This gives us a long-run group equilibrium with free entry. The same position would tend to be approached via exit of firms should sellers generally be making losses.

The normal-profit equilibrium is reached at an output short of the optimum scale of the firm, where the sloping demand curve is tangent to the cost curve. This is because, in common-sense terms, each seller continues to defend his own profit by a degree of monopoly restriction of output until additional entry has forced all of them to produce at uneconomically low rates. The aggregate output of the group of sellers, however, is restricted only to the extent that average costs are raised by the sub-optimum output rates thus effected. Monopoly thus does not restrict output seriously unless entry is effectively blocked.

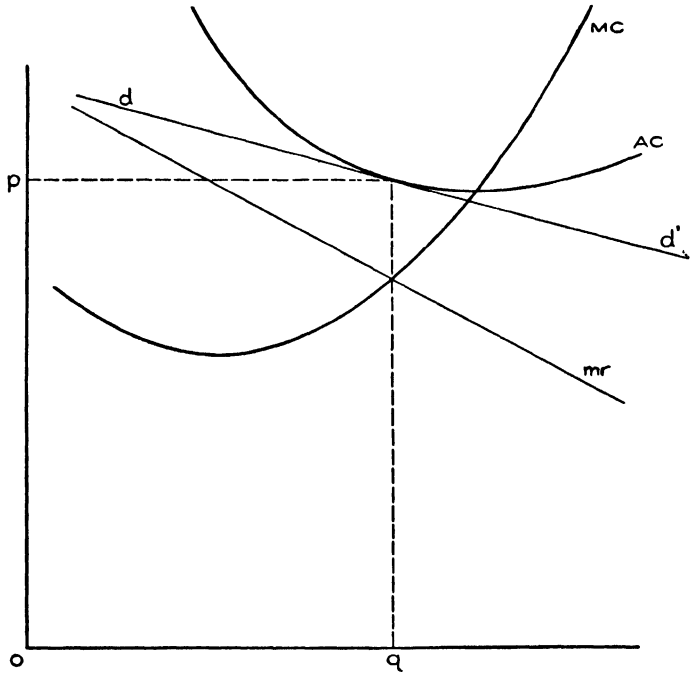

Figure 42

The character of price results in such an equilibrium is clear: (1) price is equal to average cost, allowing no excess profits; (2) price exceeds marginal cost by a relatively small amount; and (3) all firms produce at less than optimum scale. Aggregate output is restricted, however, only in very slight degree. For all firms simultaneously to reach this equilibrium would, of course, require a peculiar symmetry in the relation of their individual demands to their individual costs. In practice, such symmetry is not found, and any tentative group equilibrium would find some sellers in the normal-profit position, and others, with especially popular products or "strong" trade-marks, still earning excess profits. In effect, the "monopoly" positions of some of the sellers are ultimately rendered valueless by the pressure of entry, whereas other such positions retain some value. Tentative stability would be reached, then, when no potential entrant felt he could make normal or better profits by

entering the industry. Very large excess profits for any seller, however, are made improbable by the freedom of entry.[1]

Should product variation and selling costs be ruled out (as we have done by assumption so far), it is not clear that monopolistic competition would yield *significantly* different results than pure competition. The former would resemble the latter in that long-run excess profits would be absent or small. It would differ in that there would be a slight price–marginal-cost discrepancy and a slight departure from optimum scale whenever equilibrium was reached, reflecting an equivalent departure from minimum attainable cost of industry output. And there would be a variety of brands rather than a single homogeneous product.[2] But if we do not correct for the rotation of the earth in aiming our missiles in a snowball fight, perhaps these differences in a dynamic economy can be neglected. Having said this, it is well to remember that oligopoly and not monopolistic competition offers the dominant motif in our economy.

NONPRICE COMPETITION AMONG MANY SELLERS

Monopolistic competition is also marked by product variation and by selling costs. Let us see how this alters the picture so far developed. Like any single-firm monopoly, or any firm in differentiated oligopoly, the seller in monopolistic competition is able to vary his product—in quality, design, trade-mark, or otherwise—and to incur selling costs for advertising and otherwise promoting the sale of his product. His intention in incurring such costs is to expand his market, or to cause the demand curve for his product to shift to the right or upward. Thus he simultaneously increases his revenues and his costs. He should pursue such expenditure to the point which maximizes his profit —that is, allows the maximum difference between all costs, including selling costs, and revenues.

Each seller in monopolistic competition presumably pursues such a policy independently, considering how much he can in-

[1] For the original (and more detailed) development of this theory, see Chamberlin, *op. cit.*, Chap. 5.

[2] It may be argued that the advantage of variety to consumers compensates at least in part for the departure from minimum aggregate cost.

crease his own demand *on the assumption that his rivals' prices and products and selling outlays remain unchanged.* This is because he generally pays no direct attention to any of his rivals, their number being so large. It follows that each of the many sellers in such an industry will independently pursue product and selling policies which tend to cancel each other out, actually leaving no seller with a larger demand.[3] If this happens, such costs will tend to be incurred up to some point—possibly that where price just covers average cost for all sellers and possibly either allowing general excess profits or causing some losses to all firms. In this case, however, we recognize that profit is eliminated not by prices being forced down by entry but by costs being forced up through competitive product variation and sales promotion.

This is the solution with a given number of firms. If this selling-cost equilibrium allows excess profits, entry will be attracted—if losses, exit will occur. Exit or entry and competitive sales promotion will operate together to find an equilibrium at some point. The price results of monopolistic competition will therefore be extended to include (1) some volume of selling costs and product-variation costs, and (2) some rate of product variation or improvement. The extent to which the price-reducing force of new entry will supplant the cost-raising force of competitive sales promotion cannot be accurately predicted without a knowledge of the shape of the functional relationship between selling cost and revenue for the firm. Since entry is generally easy in monopolistic competition and more difficult in differentiated oligopoly, however, it is probable that selling costs will have freer play in the latter case. In monopolistic competition, the costs of product variation and advertising will probably be relatively low. The force of price competition acts as a potential limit on sales-promotion expenditures.[4]

[3] Except so far as total demand for all products in the industry is enhanced by sales promotion.

[4] For the original statement of the theory described in this section, see Chamberlin, *op. cit.*, Chaps. 6 and 7. The observation that the pressure of entry keeps selling costs relatively low may be justified either because the relation of selling cost to revenue for the firm is such as to elicit only moderate selling expense—which seems possible in these cases—or that the pressure of entry by

Observation of business behavior in industries which approximate conditions of monopolistic competition supports the theoretical conclusions just developed. In the manufacturing field, the ladies'-dress industry is a fair example.[5] There are many small dress manufacturers, located in New York, Los Angeles, and other metropolitan centers. Large scale offers negligible advantages in cost, since the basic technical units, after designing and pattern drafting, are the cutting shears and the sewing machine. Competition is extremely active in both price and nonprice phases. Each seller differentiates and varies his line of dresses through annual or seasonal style changes, and engages in a modicum of sales promotion through hired sales representatives, entertainment of buyers from retail stores, and so forth. Prices are ostensibly fixed in conventional "price lines"—for example, cotton dresses may be priced wholesale at $1.95, $2.95, $3.95, and so on, and rayon dresses at $8.95, $10.95, etc. Each seller produces in certain price lines. Competition takes place continually by variation of the quality, as measured in style, fabric, and workmanship, of the products offered in given price lines, and, alternatively, by shifting products of given quality from one price line to another. Since "quality" is not precisely measurable, nonprice and price competition are thus inseparably blended.

The average result of this sort of competition over time is for most sellers to drive price to the level of cost, allowing barely normal profits. Selling costs outside of product variation are quite small. It has not been ascertained whether a price–marginal-cost discrepancy occurs or whether firms operate at less than the optimum scale. But the general effects of independent price and product policies by a large number of sellers, and of unrestricted and easy entry to the industry, are evident.

The same general phenomenon is observed in groups or "industries" of retail sellers in given localities—for example, in men's haberdashery and clothing stores in any metropolis. There will be a large number of such stores, including the men's

firms with low selling costs and prices drives out those with high selling costs. In the latter case, however, the industry might tend to dynamic instability rather than to equilibrium.

[5] See Nelson and Keim, *op. cit.*, Chap. 3.

clothing departments of department stores. Nonprice competition occurs partly in terms of basic product—the quality of the clothing bought and sold—partly in terms of service, location of store, beauty and "snob-appeal" of the store, partly in terms of credit facilities offered, and partly in terms of newspaper and radio advertising. Price competition takes place through varying the regular margin or mark-up earned on clothing, and through cut-price "sales." The two sorts of competition combine in most cases to drive costs upward a bit toward price and to drive price downward toward cost. The net results are similar to those attributed to the ladies'-dress industry, with the added fact that there is some evidence of a tendency toward less than optimum-scale operations by many stores. Sales promotion costs, as distinct from the costs of performing the basic function of assembling, storing, displaying, and delivering the goods, can hardly be characterized as an excessive or large proportion of total costs.

A number of our retail distributive industries fall in the pattern just described, and more of them would if it were not for certain special phenomena. Retailing, after all, is especially suited to small-scale operations by many sellers. A first such phenomenon is the integration of retailing and manufacturing in certain industries, where the manufacturer has reached forward to acquire his own retail outlets. This has occurred in the petroleum industry in certain regions and in the optical supply business. Here the oligopolistic pattern of the manufacturing industry is imposed on retailing, even though each oligopolist may operate many retail outlets. Another factor is the growth of horizontal combinations of "chains" in drug stores, groceries, auto supplies, and the like. This may introduce enough concentration into retailing to make the market structure dominantly oligopolistic. Finally, independent retailers in the drug, grocery, liquor, appliance, and other fields have in most states secured legal interference with price competition among them, through the so-called "fair trade" laws. These have tended to lessen price competition, possibly to accentuate nonprice competition, and to attract excessive entry into the affected fields. Any detailed evaluation of competition in the distributive trades must take account of these and other complicating factors.

MONOPOLISTIC COMPETITION—A SUMMARY

It may be interesting at this point to compare the institution of monopolistic competition as a whole with other types of market organization we have considered. For markets in general three fundamental factors condition the emergent price behavior: (1) whether or not the individual seller has rivals or competitors producing an identical product; (2) whether the number of rivals producing identical or close-substitute products is none, few, or many; and (3) whether additional entry to the market is easy, difficult, or impossible. These three conditions may be labeled briefly as degree of product differentiation, degree of concentration of sellers in an industry, and degree of ease of entry. To these conditions may be added a fourth—degree of concentration of buyers.

Of these several conditions, two seem to be of dominant importance from any standpoint—number of sellers and degree of ease of entry of more sellers. Where the number of sellers is large and entry is easy (the second condition almost always accompanies the first), we generally find little net restriction of industry output below a practical maximum, little or no excess profits, reasonably low costs, freedom from excessive selling costs, and a rather flexible price over time. This is true of both pure competition and monopolistic competition, which are dominantly similar in these respects. To be sure, they have one significant difference, with respect to product differentiation. This difference tends to give rise in the case of monopolistic competition to a slightly greater number of sellers relative to a given demand, to production at slightly less than optimum scale, to a generally moderate expenditure on selling costs, and to a variety of products in an industry rather than a single standardized product. But the differences are probably not as significant as the similarities.

Where the number of sellers is few and entry is correspondingly more difficult, a basically different system of behavior is introduced. Price results, stemming from the intrinsically interdependent decisions of a few large sellers, are no longer automatically or quasi-automatically determined by impersonal

market forces, but emerge from deliberate price policies. They therefore become much more difficult to predict except by investigating a large number of considerations which condition the formation of such policies. The most significant of such considerations are the degree of ease of new entry to the industry and the degree of product differentiation. From oligopolistic industry, therefore, we get a number of significant subpatterns, depending on these and other considerations. The dominant single motif, however, is some restriction of aggregate output, some tendency toward excess profit (these tendencies being stronger as new entry is more difficult), and a tendency toward excessive selling costs where product differentiation is significant.

In general, buyers are many and small, and most industries sell to a relatively atomistic buyers' market. Where buyers are few, however, as they are for certain industries, very significant modifications in the behavior of price may take place.

With the discussion of monopolistic competition, we have concluded our investigation of pricing in various types of individual industries. It is now time to consider an economy made up of a variety of such industries, and to consider how the price system operates in this setting.

SUPPLEMENTARY READINGS

EDWARD H. CHAMBERLIN, *The Theory of Monopolistic Competition*, Chaps. 5-7.

J. E. MFADE and C. J. HITCH, *Economic Analysis and Public Policy*, Part II, Chaps. 4-6.

CHAPTER 9

THE PRICE SYSTEM FOR COMMODITIES

The preceding five chapters have considered the determination of commodity prices and outputs, the size of profits, the magnitude of selling costs, and certain related matters in a capitalist economy operating under various competitive conditions. Five principal types of selling-market structure have been considered—pure competition, monopolistic competition, pure and differentiated oligopoly, and single-firm monopoly. The analysis of such selling markets has been conducted mainly on the suppositions that industries of each sort sell to many buyers and that the underlying markets where these sellers purchase their supplies and incur their costs also have many buyers, but we have also considered the effect in either case of monopsonistic or concentrated buying.

What conclusions, or rather a priori logical deductions, can we draw from this analysis? First, we have been able to predict the price-output and related decisions of the individual firm in various industrial situations. This analysis was made by assuming a given individual-seller demand curve (or, in the case of oligopoly, a complex participation in an industry demand) appropriate to the market structure, a given set of money factor prices, and, correspondingly, some given cost curve. For other than purely competitive situations, some demand–selling-cost and other relations must also be assumed. The strategic assumptions,

however, are a given level of money demand and given cost curves. Given these assumptions, we have analyzed how a firm operating in various types of industry structure will adjust cost to demand to determine price and output, and how also it will determine product, selling cost, etc. This analysis has been modified to accommodate situations where instead of given money factor prices the firm faces a certain sloping supply curve for factors which endows it with monopsony power. As a result, we have a prediction of the "equilibrium of the firm" in pure competition, single-firm monopoly, oligopoly, and monopolistic competition both without and with concentrated buying. Such equilibria are not logically determinate with oligopoly, oligopsony, or bilateral oligopoly, although regular behavior may occur in such instances.

Second, we have analyzed the interactions of the decisions of firms within various types of industry market structures to ascertain the probable outcome with respect to price, output, and related matters for an interrelated group of firms or industry—in effect, the industry equilibrium in various competitive situations. This analysis has been made on the assumption of a given industry demand curve, to which the member firms jointly adjust and from which their interrelated individual demands are derived, and generally of given factor prices which enter into the costs of the several firms. (Account has also been taken of monopsonistic situations.) From this analysis we have gained an idea of the determination in various market structures of the price, output, selling cost, product, profits, and price flexibility—where the member firms are adjusting given cost curves to a given industry demand. Certain significant differences in the probable equilibria of monopolistic, competitive, oligopolistic, and other types of industries have thus emerged. The virtual impact of various individual industry equilibria upon the general economic welfare has also been the subject of comment.

Although the response of the single industry to a given demand and cost situation has been our primary emphasis to this point, we have also discussed at several points the interrelated behavior of a whole economy of industries in various simplified situations. We have discussed the tendency of aggregate employ-

ment, distribution of income toward profits, allocation of resources among uses, selling costs, progressiveness, and stability in economies made up entirely of purely competitive industries, entirely of single-firm monopolies, entirely of oligopolies, and so forth. This analysis has been conducted generally on the assumption of given constant flow of total money purchasing power (or aggregate demand for all products) and of an adjustable general level of money factor prices. In effect, aggregate demand has been held constant, but the general average level of prices for hired factors of production has been supposed to be freely adjustable.

A number of valid predictions may be found by arguing on these assumptions, but it is probably true that little light can thereby be cast on the determination of the aggregate level of output and employment. The coexistence of stable money purchasing power and flexible money factor prices (arbitrarily assumed) effectively begs the question of the determination of the level of employment, for if money demands for commodities, and indirectly for factors, will remain steady while factor prices seek such a level as will employ all of them, there must be full employment in the sense that all factors wishing to work will be employed. We have thus been at pains to point out that in competitive, monopolistic, oligopolistic, and monopsonistic economies, any tendency to output restriction by individual industries will not result in less than full employment *if* purchasing power is constant and factor prices freely adjustable. Under these circumstances, the impact of monopolistic or other restriction necessarily falls upon allocation, selling costs, stability, progressiveness, income distribution, etc., but not primarily on employment.

The explanation of the level of employment, and of the effect on it of various sorts of pricing, is to be found in the things which determine the ratio of the general level of factor and commodity prices to the rate of flow of money purchasing power. There is reason to believe that factor prices do not always adjust relative to money purchasing power in such wise as to permit full employment, and this regardless of whether pricing is monopolistic or competitive. Until the manner in which this

ratio is determined, and what influences it, is established, we are thus unable to assess the effects of various sorts of price-cost behavior on the level of employment. When it has been established, in Chapters 10 to 12, we may be able to see how employment is influenced by price behavior.

Anticipating this discussion, however, it may be pertinent to emphasize the various ways in which general monopolistic, oligopolistic, and monopsonistic restriction *may* lessen aggregate employment and output and result in unemployment of resources. First, if both money purchasing power and money factor prices are given and rigid, or if their ratio will not adjust in response to a change in price–average-cost margins due to monopolistic or other restriction, then a noncompetitive system will tend to employ fewer resources than a competitive system. This is because, with restriction, and with given costs, commodity prices will be higher than the competitive level and a given money purchasing power will buy fewer goods. If, then, an unrestricted or competitive system would just give full employment, a restricted or monopolized system will give less than this. A first setting in which restrictive output policies may reduce employment is therefore where there is a specific rigidity in the ratio of money factor prices to money purchasing power, and where employment would be barely adequate in the absence of restriction.

A second possible influence of monopolistic pricing on the level of employment may be via its effect on the rate of flow of money purchasing power, or on its ratio to money factor prices, *provided this ratio is subject to influence.* It may do so by affecting the distribution of income as between profit receivers and hired factors of production. At any given level of factor prices, an increase in the degree of monopolistic restriction will ordinarily mean that prices exceed average costs by larger amounts, reducing the share of all income going to wages, interest, and rents, and increasing the share going to profits.

This in turn may influence the tendency to spend and thus to create new income in two ways. First, it may influence the volume of expenditure on consumption goods. Ordinarily we should expect that an increase in the size of profits relative to

wages, etc., would tend to restrict consumption spending, since it would make income distribution more unequal, and thus increase the disposition toward saving from large incomes. On the other hand, larger profits may increase the *incentive* to business investment, thus leading to more investment spending which may partly replace, fully compensate for, or outweigh the loss of consumption spending. The net effect on money purchasing power of excess profits resulting from monopolistic restriction of output will thus depend (1) on the relative propensities toward consumptive spending of profit-receivers and of the recipients of other distributive shares, and (2) on the relation between size of profits (above the practical minimum) and the volume of investment spending. Since these magnitudes cannot be known a priori and have not been satisfactorily measured, we cannot appraise here the effect of the existing degree of monopoly restriction on the flow of money income. It may be positive, negative, or neutral. But a definite positive or negative effect on spending (relative to the level of factor prices) will have a corresponding positive or negative influence on total employment.

It should be noted, however, that if monopoly restriction is to have a favorable net effect on production and employment, it must have a sufficient *stimulus* on money income to outweigh its virtually restrictive effect on production at any given level of income and factor prices. That is, it must improve the ratio of money purchasing power to factor prices by more than enough to outweigh the increase in commodity prices relative to factor prices. Supposing that factor prices do not automatically adjust to secure full employment, but that their ratio to money purchasing power may be influenced by forces favorable to spending, then monopolistic restriction may affect employment either way, but it seems somewhat more likely to restrict it than not.

In sum, there are three alternative situations for judging the effect of general monopolistic and monopsonistic restriction on total employment:

1. The ratio of money purchasing power (Y) to money factor prices (W) is freely adjustable, because money factor

prices can and will fall relative to purchasing power so as always to secure full employment. Here the restrictions mentioned will not influence employment at all, except in so far as fewer resources choose to work with an altered income distribution.

2. The ratio of Y to W is fixed and rigid, either because both are absolutely fixed or because they must change together. Here the imposition of monopoly restriction is bound to lessen employment and output by raising the ratio of commodity prices to income and to factor prices.
3. The ratio of Y to W is not freely adjustable to secure full employment, but it may be influenced in either direction by forces affecting the propensity of income recipients to spend. Here monopolistic restriction may either increase or decrease total employment, although it has to influence spending quite favorably to cause employment to increase.

This is as much as we can say on the matter on the basis of our discussion to date. We will investigate it further in succeeding chapters.

We have also not as yet discussed the determination of the *relative* prices and shares of incomes earned by the several factors of production—land, labor, and capital—but have instead dealt with factor prices as a certain combined average price which enters into firm's costs. Until we have done this, our analysis of the price system as a whole will be incomplete. Upon the basis of our analysis so far, however, it should be appropriate to consider, assuming given purchasing power and adjustable or given factor prices, the tendencies for the real economy regarding the allocation of resources among uses, the distribution of income toward profits, the size of selling costs, efficiency, progressiveness, and stability. Certain valid conclusions concerning all of these may be drawn from our analysis to this point. What we seek is some conclusion on each of the points mentioned for an economy made up of various types of industries—in pure competition, monopolistic competition, oligopoly, and monopoly, and with elements of monopsony affecting purchases in some areas.

RESOURCE ALLOCATION

Let us first look to the matter of allocation of resources among uses, or, in other words, to the ratios in which various goods are produced. What is the ratio of beef output to auto output to the output of aluminum coffeepots, and is it the best ratio from the standpoint of buyer satisfaction? This analysis is most easily conducted on the simple assumptions (previously adopted) that a certain aggregate volume of resources is employed in all uses, that there are a certain number of goods to produce, and that all the price system does is to fix the ratio in which the various goods are produced. We assume that there is a given *level of employment of resources* and that the problem is simply how these resources shall be allocated among a predetermined range of uses or lines of production.

When the problem is thus simplified, it can be shown that buyers will get the most satisfaction from a given amount of employed resources if those resources are so allocated that a certain uniform balance is struck between price and cost in every line of production. Specifically, their satisfaction will be greatest if the goods are produced in just such proportions that for every good the ratio of price to marginal cost is the same.[1] (Marginal cost here will refer throughout to a measure of marginal cost in terms of real resources used—*e.g.*, to the money value of marginal real cost.) If this condition is met, the last unit of real cost in every line nets an equal amount of satisfaction (as measured by price), and no shift in allocation could then enhance total satisfaction.[2]

We have previously shown that if all of the given range of goods were produced under conditions of pure competition, every price would tend toward equality with its marginal costs, and allocation would then be demonstrably ideal. We have also indicated that if certain goods were monopolized, so that mar-

[1] But see page 165, note 12, for qualification of this point.

[2] This argument postulates the intercommensurability of different buyers' satisfactions—*i.e.*, supposes that if Brown pays $1 for a necktie, it represents the same satisfaction that Smith gets from a penknife bought for $1. Some such assumption is essential to any concept of aggregate satisfaction.

ginal cost was held short of price, whereas others were produced so that marginal cost equaled price, allocation would not be ideal, since a balanced ratio between price and marginal cost would not obtain and could be effected only by shifting resources toward the production of the monopolized goods and away from the other lines. A further proposition can be supported to the effect that if all industries were monopolized, but "in equal degree," so that marginal cost everywhere was short of price in the same ratio, allocation would also be ideal.[3] Nonideal allocation results when the ratio of price to marginal cost is different for different industries.

With those general guides, what can be said of allocation at a given level of employment in the economy we have? In this economy, certain markets are purely competitive, with prices tending to equal marginal cost; some are in monopolistic competition, with presumably slight price–marginal-cost discrepancies; a few are single-firm monopolies, with larger discrepancies of this sort, except where public regulation intervenes; at least a majority are oligopolistic, and in these the relations of price to marginal cost are various and unpredictable. Monopsonistic elements introduce further discrepancies. If one good came directly to consumers from each market, it would be extremely improbable that any close approximation to ideal allocation would result. With each final good passing through and affected by a vertical sequence of markets, from raw materials to retail distribution, and with different markets in the sequence having different structures and giving rise to different price-cost relationships, the allocation picture is further confused.

It is in fact quite impossible to say a priori just what sort of allocation our system gives us, or how far it departs from the ideal, except to note that very close approximation to the ideal is highly improbable. It is quite possible that in a significant degree we get relatively too few of certain goods and too many of others. At the same time, it must be pointed out that in a free-enterprise system, where any restrictive monopoly is subject to the interference of competition or of public authority if it is restrictive beyond a certain point, there are definite limits

[3] See page 165, note 12, for qualification of this rule.

to the distortion of allocation. The discrepancies from the ideal at any time are not likely to be huge.

PRODUCTIVE EFFICIENCY AND INCOME DISTRIBUTION

In preceding chapters we have already referred to the general impact of the commodity-pricing system on efficiency and on income distribution. These matters may therefore be dealt with here quite briefly. Efficiency is affected in the pricing process so far as firms are led thereby to produce at scales of operation or rates of utilization other than the attainable ideals.[4] It would appear that in certain sectors of our economy, especially where excessive investment is attracted into oligopolistic markets or markets in monopolistic competition, and is maintained there by monopoly pricing tactics, productive efficiency is adversely affected. The effect of monopolistic pricing in making income distribution more unequal is so obvious as to require no added comment, other than that the resulting current distortions of income distribution, and the accumulated effect on the distribution of wealth, are quite significant from the standpoint of social welfare.

THE SIZE OF SELLING COSTS

Another thing affected by the operation of business competition is the allocation of resources between the production of goods and the selling and distribution of goods. We have indicated that in oligopolistic industries which sell consumer's goods, and to a lesser extent in industries in monopolistic competition, there is a tendency to incur selling costs which are a very significant proportion of total costs, and also to expend significant amounts on product improvement and variation. This tendency in varying degree affects virtually the whole range of consumer's goods in our economy. When the tendencies noted in many individual industries are aggregated, we have a major characteristic of the whole economy. In effect, the sort of business rivalry common in most consumer's-goods markets has led

[4] Cf. pages 120, 153, and 192 above.

to the diversion of an important fraction of our employable resources to "selling" activities of one sort or another. The net return which buyers obtain from this activity in variety of products, quality of output, convenience, and incidental entertainment probably does not offset the virtual loss in quantity of basic production. The system may therefore be criticized not for incurring selling costs at all, since some are needed, but for incurring them beyond a justifiable limit.

PROGRESSIVENESS IN THE MODERN ECONOMY

Two additional aspects of the performance of an economy are its progressiveness and its stability. Progressiveness, which may be measured as the rate at which an economy expands its aggregate or per-capita output over time, is influenced by many factors, of which the character of commodity-market structures is only one. The question of whether business operations within the existing market frameworks are as conducive to progress as they would be under some other feasible framework is, therefore, perhaps unduly narrow. On the record, the capitalist economy, just as it has been and is, has been extremely progressive both in growth of productivity and in introduction of new techniques and products. This historical progressiveness is perhaps its principal claim to eminence as a system of economic organization. Upon this background, the main question raised to this point is whether the monopolistic and quasi-monopolistic tendencies evident in the present-day economy are favorable to unfavorable to continued progress.

There are a number of possible answers to this question. One would hold that the great progress registered by capitalism was the result of free and active competition, and that the increasing concentration of markets into tight oligopolies of firms, afraid to compete actively, is inimical to progress via innovation and forward-looking investment. Another answer would maintain that a highly concentrated and quasi-monopolistic business organization is best equipped to accomplish the marvels of research and engineering which are essential to continued progress. Still a third would hold that the profits of monopoly are a necessary lure to innovation, and always have been, and that a pattern of

monopolistic restriction is an intrinsic and, in the net, desirable aspect of capitalism.

From the analysis we have developed in preceding chapters, only a certain amount of light can be shed on this rather controverted issue:

1. Although it is probably true that the possibility of securing a superior earning position and protecting it in the form of a monopoly is a lure to innovation and progress, the continued defense of old monopoly positions may restrict progress. Further, when concentration proceeds to the point where one firm or closely knit group at the same time benefits from an old monopoly position *and* has the principal chance to innovate and replace the old monopoly with a new one, the probability of innovation is decidedly smaller than where the old monopolist and the potential innovator are distinctly different people. This argues that *very* highly concentrated market structures may be relatively inimical to progress.

2. The typically oligopolistic market structure of the American economy, however, ordinarily has not as yet proceeded far enough toward collusive monopoly seriously to reduce competitive innovation. In most of our oligopolies there is very active nonprice competition, and this tends to promote a fairly rapid rate of progress in adopting new techniques and introducing new products. Moderately concentrated oligopolies may well turn out a more rapid rate of progress in technique and product than any other sort of market organization. (It must be remembered, however, that a suppressive tendency on employment or a distortion of income distribution may counterbalance this tendency to progress.)

3. Further development of market structures toward single-firm monopolies would probably be inimical to progress.

4. Within the general framework of modern market structures, less restrictive price and output policies would be consistent with the maintenance of strong progressive tendencies.

ECONOMIC STABILITY

The effect of the system of commodity pricing on the stability of the economy or, conversely, on its susceptibility to

fluctuations of income and employment, has already been referred to at several junctures. As a general background for any discussion along this line we must recognize that the capitalist economy will experience fluctuations in income and employment pretty much regardless of the relations of commodity prices to costs. Observation and theory both support this conclusion. We must also recognize that such fluctuations in income will generally be accompanied by connected fluctuations in wages and other factor prices, and that such fluctuations in the costs of producing commodities will probably have more effect on the course of the "business cycles" than will movements in price-cost relationships. Finally, therefore, the influence of such price-cost relations on stability will be only moderate.

The responses of price-cost relationships and of commodity prices to fluctuations in income form a rather complicated pattern for our whole economy. As money income expands with upswings or movements toward prosperity, the demand curves for most products shift upward and to the right, tending to elicit both larger outputs and higher prices. At the same time factor prices (wages, interest, rents) also rise, though often lagging the rise in demands, causing the cost curves of firms to shift upward. This further accentuates the tendency of prices to rise, but tempers the rise of output somewhat. As money income declines with downswings or movements toward depression, the reverse of these movements takes place: demand curves shift down to the left, tending to reduce prices and outputs, and the fall in factor prices shifts cost curves downward, accentuating the decline in price but tempering the decline in output. So much for the general mechanics of price-output adjustments in response to fluctuations in income or purchasing power under competitive or quasi-competitive conditions. Where factor-price determination takes place largely in bilateral monopoly situations, less definite predictions can be made.

It is also important to note that as income fluctuations take place, factor prices are for a variety of reasons *relatively* inflexible or sticky, so that they do not fluctuate as rapidly or as widely as income does, and do not fully compensate for these fluctuations. The costs of producing commodities are therefore relatively inflexible, and practically regardless of how com-

modity prices adjust to these costs, the basic inflexibility of factor prices makes it likely that income fluctuations will cause wide fluctuations in employment and output.

The issue with respect to commodity-price movements is, therefore, how commodity prices adjust to shifting demands and shifting costs over the course of the business cycle, and what effect this adjustment has on the nature and impact of the basic income fluctuation. The salient aspect of these adjustments in the modern economy is that goods prices behave in various ways because of differences in market structures and price-determining situations. We have a limited sector of the economy in some approximation to pure competition, including a number of agricultural products and a few basic manufactures, although government interference with price and output in this sector is becoming the rule. A large sector of our extractive, processing, and manufacturing industry forms an oligopoly area in the economy, where pricing follows certain patterns peculiar to this category. Some manufacturing industries and many distributive industries are in some approximation to monopolistic competition. Finally, most public-utility industries are dominated by local single-firm monopolies, whose prices, however, are determined by public regulatory bodies.

The tendency of price in unregulated pure competition is to respond very actively and quite automatically to fluctuations in demand and cost. As we saw in our analysis of purely competitive markets, a moderate shift in demand in a purely competitive industry will ordinarily produce a sizable shift in price even if the level of costs (factor prices) is unchanged. When costs also shift in the same direction as demand, the shift in price is even larger. In the purely competitive sector, therefore, we tend in the absence of regulation to get very flexible prices over a cycle of income and, correspondingly, relatively stable outputs. Price movements compensate for much of the shift in demand; therefore output is much less affected. This movement takes place independently of the control of any sellers, who cannot influence the movement of market prices. Such price behavior tends to minimize the fluctuation in output. But it is likely to maximize the corresponding fluctuations in the short-

period profits of the sellers involved, and therefore to make the whole situation rather unpopular with them.

Further, it is possible that the very flexibility of price will accentuate the tendency of buyers to anticipate price fluctuations by buying in advance during upswings and withholding purchases during downswings, and that this may cause such upswings and downswings in income to proceed more rapidly and possibly to greater extremes. So much for the impact of unregulated purely competitive pricing in a fluctuating economy.

Within the oligopoly and unregulated-monopoly sector of the economy, there are many variants in price behavior. If there is a central tendency, however, it is toward relative inflexibility of prices in the face of fluctuating income. The rationale of inflexible price policies in monopoly and oligopoly has been developed in preceding chapters. We have also emphasized the *ability* of sellers in such markets to control prices deliberately in spite of shifting demands. Statistical studies indicate that in fact prices in the industrial sector have changed less frequently and by smaller percentage amounts than prices in the purely competitive or "market-controlled" sphere.

The impact of such pricing on the economy is fairly evident. Prices respond less readily to shifts in demand and in cost, and therefore output and employment fluctuate more. By the same token, the short-run profits of sellers fluctuate less than they do in pure competition. The impact on output and employment of given fluctuations of income are therefore accentuated. At the same time, the tendency of buyers' speculation on price movements to accentuate income movements may be considerably less than in pure competition.

Prices in regulated single-firm monopolies generally tend to be quite inflexible because of the rigidities implicit in such regulation, and they thus augment the tendencies of the oligopoly sector. Prices in monopolistic competition are ordinarily fairly flexible over time, and may fall in with the purely competitive sector. We thus tend to have, in a broad sense, a flexible-price sector and an inflexible-price sector in the economy, each with its own virtual impact on the course and effect of fluctuations.

When these sectors are combined and operate together, as they do, certain additional phenomena emerge. First, the flexible-

price sector of the economy is considerable enough that the general response of prices to income fluctuations is sufficient to set up cumulative movements in the economy—to cause buyers to speculate on price changes enough to accentuate upswings and downswings in income. The whole price system is *not rigid enough* to forestall cumulative movements in income based on speculation. Therefore such considerable partial rigidity as there is only accentuates the response of output and employment to income fluctuations, without adequately dampening the basic fluctuations of income. The mixed system is therefore particularly unstable, and susceptible to maximum swings of employment and output.

A second result of the mixture of flexible and inflexible prices is that the net incomes of sellers in the flexible-price sector—mainly farmers—tend to fluctuate much more than those of sellers in the concentrated industrial sectors. Farmers therefore tend to be very poor indeed in business depressions, even though they may fare exceptionally well in prosperity. Their depression plight is accentuated by the fact that although their incomes fall off steeply in depression, the manufactured goods they must buy are inflexibly priced. This situation may lead to serious social problems, to regional economic distress, to banking and mortgage difficulties, and so on. It was this special distress in the agricultural sphere which led certain economists to emphasize the price rigidity problem, and to propose government crop control and agricultural price support as a means of improving the financial welfare of farmers. As a result we are currently tending toward artificially pegged farm prices and an entire economy of relatively inflexible prices. The merits of such a system may be debated.[5]

The preceding tentative conclusions concerning the function of the commodity price system are, of course, subject to modification and extension after we have taken into account the behavior of wages, interest, and rents. We therefore turn next to the analysis of factor-price determination.

[5] For a discussion of the effects of flexible and inflexible prices, see Alvin H. Hansen, *Fiscal Policy and Business Cycles* (New York: W. W. Norton & Company, 1941), Chap. 15.

SUPPLEMENTARY READINGS

DONALD H. WALLACE, "Industrial Markets and Public Policy," *Publu Policy Yearbook*, Harvard Graduate School of Public Administration, Cambridge, Mass., 1940.

BEN W. LEWIS AND OTHERS, *Economic Standards of Government Price Control*, Temporary National Economic Committee, Monograph No. 32, Washington, 1941.

CHAPTER 10

THE DISTRIBUTION OF INCOME

THE PROBLEMS OF EMPLOYMENT AND INCOME DISTRIBUTION

In preceding chapters we have analyzed the determination of various commodity prices and outputs, and of the allocation of resources among uses, as it occurs in a free-enterprise economy. This theory of commodity pricing has been deduced, however, subject to one principal limiting assumption—that there is for the economy a given constant flow of money purchasing power, to which both commodity and factor prices may adjust without influencing the flow itself. We have thus seen what would be the equilibrium tendency of relative commodity prices and outputs if purchasing power were given and constant and if factor prices (which make up costs) were free to adjust to such a given volume of buying power. In conducting this analysis, moreover, we have so far failed to distinguish among the particular prices of different factors of production—labor, land, and capital—speaking instead essentially of a certain average or composite price of such factors as it might adjust to a certain money purchasing power. Although we have observed certain things about the determination of the average money price of all productive factors in various situations, we have not heretofore analyzed the determination of the relative (comparative) prices and income shares of the different factors of production.

For certain purposes theorizing on this restricted plane is a valid and useful procedure. It allows us to examine the tend-

encies present in the determination of the outputs and price-cost relationships for commodities sold in various market situations, and thus in the allocation of productive factors among the production of various goods. It also reveals some tendencies present in the determination of the relative shares of income going to hired factors as a group and to profits as the average price of hired factors adjusts to any going level of money purchasing power.

But by the nature of the assumptions made so far we have avoided two essential problems: the determination of (1) the relative prices and income shares paid to different factors of production—that is, of the "functional" distribution of the income paid out as the cost of producing commodities; and (2) the relation between the volume of purchasing power and the average level of factor prices, which evidently helps to fix the level of employment. It is clear that we have not dealt with the first question. And we have so far begged the second by assuming purchasing power to be a constant magnitude to which factor prices are free to adjust. We must now relax our restrictive assumptions and attempt to build a single, unified explanation of the simultaneous determination of commodity prices and outputs, wages, interest rates, rents, profits, and the level of employment of the several productive factors.

PRODUCTIVE FACTORS AND DISTRIBUTIVE SHARES

Given the broad problem, we may digress momentarily to consider certain fundamental terms and ideas useful in the succeeding analysis. Business enterprises in general function by acquiring certain basic productive services, either from human beings or from the earth's resources, and by combining them into useful products. All production must draw entirely and exclusively on such basic services. One firm may buy goods (semi-processed basic resources) from another, but in so doing it is merely acquiring these basic services indirectly. The "factors of production" include all such services, or the sources which provide them. "Factor price determination" refers to the pricing of such services, whereas commodity price determination refers to the pricing of the goods made by combining them.

Basic productive services may be conveniently subdivided on the basis of origin. A common and useful subdivision recognizes two productive factors—"labor" and "land." "Labor" in this sense is used to refer to human beings in general, or to whatever productive services they provide. "Land" is used to refer to natural resources in general, or the services they provide, and would thus include agricultural land, factory sites, urban residential and commercial land, forest land, coal mines, oil wells, deposits of uranium ore, etc. Each of these "factors" may be subdivided into as many grades or types as may be convenient for analysis—labor into various geographical and occupational groups, and land into such subcategories as those just indicated. The payments to or prices of labor are generally referred to as "wages," those of land as "rents." These terms are thus given somewhat broader definitions than is common in nontechnical discussion.

Are labor and land thus defined the only factors of production? A third candidate for inclusion is often suggested, namely, "*capital*"; this is also often held to be a "factor of production." The sense in which capital may be so regarded, however, requires careful definition. Capital in the sense of capital goods—factory buildings, machinery, inventories, or any good to be used in further production—evidently represents a special class of commodities, produced for use in the production of further commodities or services. In an ultimate sense, capital goods may thus be classed with other commodities, as the output or embodiment of the basic services of land and labor. As such, they would not claim status as an additional factor of production.

At any given time in a developed industrial community, however, there is on hand a large body of previously produced capital goods, which on the average are quite durable and will contribute for some time to further production. From any current standpoint, this existing stock of capital goods represents a body of fixed factors, akin to land, upon the services of which production may draw.

Any short-run analysis will recognize existing capital goods as a third factor of production, earning (potentially) "quasi rents." As we contemplate longer periods for analysis, however, it must be recognized that existing capital goods will wear out

and have to be replaced, and that new or additional capital goods may be added. For such longer periods, capital goods revert to the status of produced commodities, or, in effect, of the output or embodiment of the services of land, of labor, and of the old capital stock on hand at the beginning point of analysis. (The last represents for this purpose a species of exhaustible "land.") Capital goods, aside from the existing capital stock which is always given as a supplement to land and resources at any current moment, thus do not fully qualify in the long run as a third factor of production.

But the initial acquisition and the retention (by replacement) of capital goods does require the use of funds or money, to be "invested in" such goods—that is, paid for them without the immediate yield of consumption satisfaction to those who make the payment. The productive services which go into making capital goods must be paid for, and in advance of the time when the capital goods provide services in further production; purchasing power must be "tied up" in them for an interval during which those who supplied the purchasing power receive no direct benefit in consumer goods. It may thus be argued that "money capital," or investable funds, constitutes a third productive factor, which is used in conjunction with other factors in the process of production to facilitate the production and use of capital goods. Whatever we call them, it is evident that investable funds are required, and also that the supply of such funds is a source of earnings to their suppliers, in the form of *interest payments*. These arise as the suppliers of funds pay less for capital goods than the goods are expected to earn over time as they are used, or, in effect, as they pay the cost of producing capital goods and receive an income stream the gross sum of which exceeds this cost.

Within a capitalist system, the interest paid for invested money is thus a third distributive share, in addition to wages and rents. It is paid for the services of invested money, *and it is "earned"* (as a part of quasi rents) *by the capital goods in which the funds are invested.* The basic additional factor may thus be regarded as investable funds, and the basic service that of investing these funds to finance acquisition and retention of capital goods. The provision of this service is in turn reflected,

however, in the continued production and use of actual capital goods, which in use provide direct productive services which are rewarded sufficiently to pay not only their cost of production but also the interest on the funds the investment of which made their acquisition possible. The use of investable funds, to which interest payments are made, is associated with the production and existence of a special class of commodities—capital goods—which are produced for use in production in conjunction with labor and land, and which directly earn the income from which interest is paid.

It may be noted, however, that investable funds are advanced not only for investment in capital goods, but also for such things as consumer loans, or for the purchase of any asset, including land, which may yield a future income stream. Such funds will also earn interest in these pursuits, so far as less will be paid for any future income stream than the gross sum of all payments in this stream. Thus the use of investable funds and the earning of interest arises not only in connection with the use of produced capital goods, but generally in connection with the purchase of the right to any future or nonimmediate series of incomes. Interest is mentioned primarily in connection with capital goods because they are the principal source of such series of delayed or future returns. But interest also is earned if a future series of rents is bought,[1] and, if labor power could be sold, could be earned in buying the rights to the future wages of labor.

It will be noted that in addition to wages, rent, and interest, there is apparently a fourth distributive share—namely, the *profits* going to enterprise. Correspondingly, it has sometimes been the fashion to suggest that we should recognize a fourth productive factor, called *enterprise* or *entrepreneurship*. Most of the functions performed by the ownership-management group in the modern enterprise, however, consist in the supply of the other productive factors—particularly the *labor* of management and administration and the money *capital* for invest-

[1] Land, or any other nonwasting or irreplaceable asset, has no "cost of production" above which it can be said to earn an interest return. But it will tend to have a market price which is sufficiently below the undiscounted sum of its future returns that an interest return is allowed on the funds paid by its purchaser.

nent in the concern. To explain the earnings of such labor and capital it is unnecessary to recognize a separate factor of production; it is only logically consistent to attribute such earnings to their true sources. After deduction of the costs of labor and capital supplied by the enterprise group, nevertheless, a residual may remain which also goes to the owner-manager group. This share may be meaningfully referred to as true *profits*, a fourth distributive share. But it is pointless for most purposes to refer to a fourth productive factor corresponding to profits. We may say, in effect, that enterprises use labor, land, and capital, disbursing distributive shares of wages, rent, and interest; that they sell the resultant product for a revenue; and that any residual difference between the sales revenue and the total of the payments mentioned goes as a true profit to enterprise. Profits are thus logically defined as *net of* or *in addition to* all wages (including wages of management), all interest (including interest on owners' investment), and all rents attributable to natural resources employed. Under general equilibrium conditions, they should represent mainly a return to artificial or noncompetitive restriction of output, and no "productive" function is rendered in return for them. Under conditions of dynamic change, however, the profit residual may be enhanced by enterprise activities which might be viewed as a special productive function corresponding to the reward. This will be discussed in Chapter 14.

In sum, the system of business enterprises which operate in selling commodities also operate by buying the services of productive factors, which may be classified as labor, land, and "capital." Enterprises make money payments for these services, which, as wages, rent, and interest, constitute (together with residual profits) the flow of money income upon which people depend for a living. One reason for distinguishing among the incomes paid for various sorts of productive services is found in differences in their character. Labor yields human services, land the services of inanimate natural resources, and "capital" the services of invested money. Another reason is found in the difference in the relation of the human recipient of the income to the service for which it is paid. Wages are ordinarily payments to the individual for his own labor—for expended effort of some

sort. Rents, as received for the services of natural resources, are not paid as rewards for effort expended by the recipient but accrue by virtue of the *ownership* of the resources in question. The institution of rent payments as incomes to individuals thus emerges from the character of *property* in natural resources. Interest similarly is a payment not for human service (unless for the "abstinence" of the supplier when he is an individual supplying his savings) but for the release of money funds which the individual or institution in question has the power to release. It is thus again a payment to ownership or to institutional position. Profits, as a possible (but not inevitable) residual left to the enterprise or to those who own or control it, also evidently have a source in ownership or property. It is in part because of the important distinction between human effort on the one hand, and property or institutional position on the other, as sources of income that it is important to trace the determination of the relative shares of income going to labor and to land, capital, and enterprise.

A GENERAL ANALYSIS OF THE DISTRIBUTIVE PROBLEM

Let us now return to economic behavior, with particular reference to the determination of wage, interest, rent, and profit incomes. A first step is to see precisely how the ensuing steps of analysis fit into what has gone before.

So far, we have been concerned with this question: Suppose that there is for the economy a given constant flow of money purchasing power seeking commodities and a given state of buyers' choices among commodities, which together provide a given family of interrelated money demand curves for all goods. Suppose also that the average money price of employed factors of production is free to adjust to this purchasing power. *Then,* how will the prices and outputs of all commodities be determined, and also the relation of their prices to their average and marginal costs—in effect, how will commodity prices and outputs, and the average factor price, *adjust* to a given aggregate flow of money purchasing power? This inquiry thus concerned the determination of the relative outputs of various goods, and of the relation of commodity prices to costs. In deal-

ing with it, we have also cast light directly on the determination of relative efficiency in the production of various goods and on the allocation of resources between production and selling. Some indirect light has also been thrown on the tendencies to progressiveness and stability inherent in a free-enterprise economy.

But having said so much, we are still far from a complete and unified explanation of how an economy works. There are a number of very important matters still to be considered. So far we have assumed buyers' purchasing power to be given in spite of adjustments in factor prices, so that the two would not be interdependent. They are interdependent in fact, however, since the payments made to factors are the principal source and determinant of the flow of purchasing power which constitutes the demand for commodities. In incurring costs of production, enterprises pay out money incomes in wages, rent, interest. These incomes, together with profits, constitute the flow of money payments which can be expended for the goods produced by hiring the productive factors. The aggregate of money income payments, or costs, flows through the economy to become the aggregate of money demands for the outputs produced. In effect, there is a *circular flow* in economic activity, with a complete circular interdependence among all sectors of the economy.

This circular flow proceeds, in reciprocating fashion, simultaneously in two directions. In Figure 43, we may conceive of the principal "way stations" in this perpetual flow as (1) enterprises, (2) recipients of income from enterprises (which in general supply enterprises with the services of factors of production), including profit receivers, and (3) buyers of the outputs of enterprises. Categories 2 and 3, however, are actually identical, since income recipients are the buyers of outputs. Within this model, we may view a flow of real services and goods as proceeding in a clockwise direction. The services of factors of production flow from *A* to *B*, as labor, land, and capital are employed. At *B* they are combined and emerge in the form of outputs of goods and services, and thence flow back to buyers of these outputs (who are also the suppliers of productive services), thus completing the circle. Reciprocating with this real flow and proceeding in the opposite direction is a flow of money payments. Thus counterclockwise from *B* to *A* runs a flow of

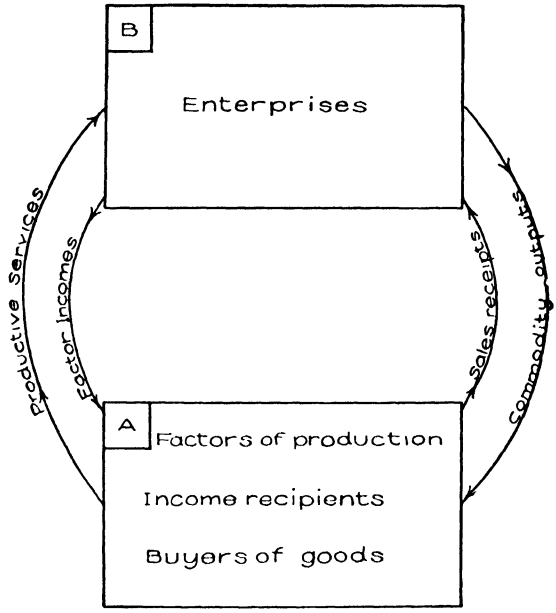

Figure 43

wage, interest, and rent payments, constituting remuneration for hired productive services, and of profits, which are the residual rewards (if any) left to enterprise. From *A* counterclockwise to *B* the flow of money continues from these income recipients back to enterprises as payment for the outputs which are being bought. This flow constitutes the sales receipts of enterprises and completes the circular flow of payments in the economy.

One facet of the circular interdependence of the economy is quite apparent. The size of the flow of real outputs to buyers depends directly upon the flow of productive services from factors to enterprise, and, conversely, the size of the flow of money sales receipts—of aggregate money demands for goods—depends directly upon the flow of money payments made by enterprises to factors. The second aspect of this proposition may require more emphasis than the first. The first, after all, refers to the obvious fact that the rate of output of the economy depends on the rate of work or employment of resources. The harder or longer we work, the more we produce, and this would be as

true in a primitive economy (without enterprise, hiring, and money exchange for goods) as it is in the economy we have. The second point, however, is exclusively pertinent to a specialized economy depending on monetary exchange, where the services of productive factors are bought for money, and where this money is in turn used to buy the output of the productive factors. In such an exchange system, there are two aggregate quantities (or rates of flow) which deserve attention: (1) the aggregate cost (including profits) of all goods produced by enterprises, which is also the aggregate of money income payments and may be referred to as the *aggregate supply price* of all output, and (2) the aggregate receipts of all enterprises, received in money payment for sale of their outputs, which may be referred to as the *aggregate demand price* of all output. The *aggregate supply price* and the *aggregate demand price* of all output are obviously closely interdependent. Each is the primary source of the other, in a flow through time. The amount of wage and related payments which enterprises make is the leading determinant of the total money demand for goods, and the total money demand for goods is the leading determinant of the volume of wage and similar payments.

But the aggregate supply and demand prices are not necessarily identical. As the circular flow proceeds through time, it is quite possible for the demand price to exceed the supply price, or for the supply price to exceed the demand price. This is because income recipients as a group may decide to spend less than they receive, "hoarding" money or increasing their liquid balances, or to spend more than they receive, "dishoarding" money or decreasing their liquid balances. If they do the former, the aggregate flow of money payments must diminish; if they do the latter, the flow of payments will increase. Only when income recipients are disposed to spend just what they receive will the flow remain constant. If stability in the flow of payments is to result, there must be an attainable rate of flow of money payments for which the receipts and expenditures of income recipients will be equal.

One evident task of theoretical analysis is to explain what determines the relation of income to expenditure and under what conditions (1) stability and (2) transient or progressive insta-

bility of the money flow may be expected. This is an important issue because the size of the money flow may influence the average level of commodity prices, the level of factor prices, the level of aggregate output, and the level of employment. Output and employment may be said to depend jointly upon the size of the money flow and the levels of factor and commodity prices. Prices of all sorts, outputs, employment, and the money flow mutually interact to determine some net outcome for the economy. If we are to explain how the economy works, we must deal with this interaction. We must explain the behavior of the circular flow of money payments (its size, its movements, and its equilibrium), the behavior of the counterflow of employment and output, and the determination of the intermediate ratios between these flows—the average money prices of commodities and of factors. This is a first principal task of a theory of income.

A second general problem concerns the relative shares of a given aggregate income which are paid to the various factors of production. In hiring productive factors and making payments to them, enterprises are, in effect, determining the distribution of income among various sectors of the economy. It is therefore important to inquire how the relative size of these shares is determined, and what particular forces affect this determination. The shares of income received by various factors obviously depend in a degree upon the general level of money payments to all factors, the level of employment and output, and the level of money prices for factors and commodities, and any valid explanation should recognize this interrelationship.

The analytical task as a whole is that of explaining the rate and composition of the circular flow of economic activity at all stages. To this point we have considered only one arc of this flow—extending from the expressed money demands of buyers for commodities through commodity outputs and prices and to (but not including) the payment of incomes to the productive factors. It remains to complete the circle, from the payments of incomes to factors and through the expression of money demands for commodities. In treating this latter arc of the circle, we must consider the size or width of the money income flow, the size of the reciprocating flow of employment, and the proportions of the income and employment flows going to and

coming from various factors of production. We must also thus inferentially consider the size of the flows of expenditure and of aggregate output.

PURCHASE OF PRODUCTIVE FACTORS BY THE FIRM—PURE COMPETITION

A convenient first step in this analysis is to follow the flow of money payments one step further, from the enterprise, as it incurs its cost of production, to the various productive factors.

Such payments are made within a variety of market structures. Certain generally applicable principles may be illustrated, however, for the simple situation of a purely competitive firm which buys factors in purely competitive factor markets. In effect, two sorts of market situations may condition the firm's purchase of productive factors. First, it is significant whether the enterprise which purchases productive factors is selling its output under competitive or monopolistic conditions, since this implicitly influences its demand for the factors. We assume provisionally that the enterprise is selling its product in a purely competitive market. Second, it is significant whether the markets in which factors are purchased are competitive—whether there are many or few buyers (*i.e.*, enterprises) bidding for factors, and whether there are many or few sellers of factors (many independent laborers, for example, or a few labor unions). We assume provisionally that each factor market is purely competitive—that there are many small buyers and many small sellers of each factor of production and also that all units of any factor are homogeneous. No enterprise buying a factor, that is, has any degree of monopsony in buying; it must accept the going price of any factor as outside its influence. And no seller of a factor has any degree of monopoly in selling; it is thus limited to deciding how much to supply at any going market price. The seller of a factor in such a situation stands in the same relation to the price of the services he offers as the seller of wheat stands to the price of wheat—he can take it or leave it, but he cannot himself change it. The student will at once recognize that many actual factor markets must have strong elements of monopsony and monopoly. Attention to purely competitive markets thus is

useful mainly for expository purposes, and not as a very close approximation to real life. For purposes of an initial simplified exposition in this chapter, we will further assume that each productive factor is equivalent to some basic inanimate commodity, available in limited quantity without being produced, and that the supply of each factor is quite independent of that of every other. (This assumption will be modified in the following chapter.)

In the competitive situation just described, let us first investigate the activities, in the long run,[2] of the individual firm *buying factors.* These will stem from the firm's demand for such factors and from the conditions of supply for each of them. The firm's demand for factors is essentially derived from the demand for its output; for the firm with a purely competitive selling market, the latter is a perfectly elastic horizontal line at the going market price for the output. This market price serves as the beginning point for the firm's calculation of its demand for factors.

If the factor markets are purely competitive, the supply of each factor to the firm is also perfectly elastic at some going price. Since the price of labor, for example, is beyond influence by any seller or buyer thereof, the supply curve which represents the amounts of labor any firm can buy at various prices is simply a horizontal line at the going wage rate. (See Figure 44.) The firm in effect can take as little or as much labor as it wishes at the going wage rate, but cannot lower and would not wish to raise this wage. The same is true in turn of each factor market; in each, the firm is faced with a given fixed price.

Suppose now that the firm is buying three factors—land, labor, and capital, or, let us say, factors A, B, and C, each with a given money price. It is also faced in its selling market with a given money price for its output. Its only decision with respect to the factor markets, then, is to decide how much of each factor to buy. How is this decision made? The complexity of the choice

[2] Attention is centered here on the firm's long-run adjustments in purchasing factors—on its adjustments, that is, over periods long enough that the quantities of all hired factors can be freely varied. Any special analysis of short-period adjustments is here neglected, but it may be easily worked out.

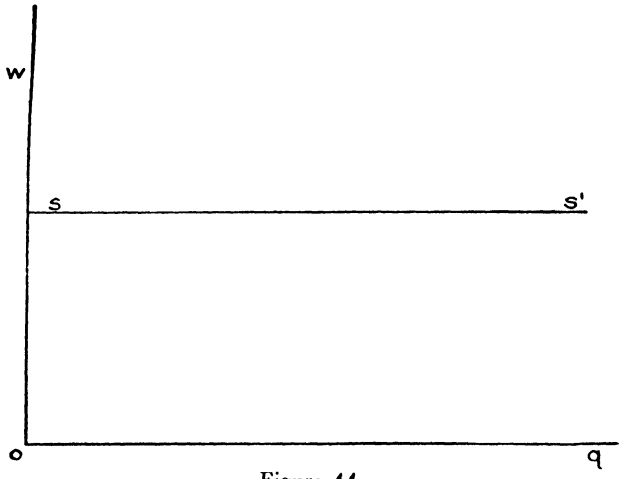

Figure 44

will depend upon whether the proportions of the various factors are fixed or variable—upon whether or not the firm may acquire the factors in proportions which vary with variations in their relative prices. We will therefore successively consider (1) purchase of factors in predetermined fixed proportions, and (2) purchase of factors in potentially varying proportions.

In the first situation, the firm must observe certain technologically fixed ratios or coefficients among the several factors; for example, for each 5 units of factor *A*, it must employ 3 units of *B*, and 1 unit of *C*. (This might correspond to 50 laborers, $30,000 invested in plant, and 1 city lot of land.) Increases in output can be accomplished only by equal percentage increases in the amounts of all factors employed. In effect, the firm cannot substitute one factor for another in any degree, and as a result the various factors do not compete with each other. Output must be increased by fixed "doses" of combined factors—in the example in question, always in a 5-3-1 ratio. In addition, the price of each factor is given to the firm, let us say at $5 per unit for *A*, $8 for *B*, and $10 for *C*. Each additional dose of factors would thus cost the firm a constant amount—in this case ($25 + $24 + $10) = $59. These payments constitute the costs of production of the firm. The average cost per unit of output

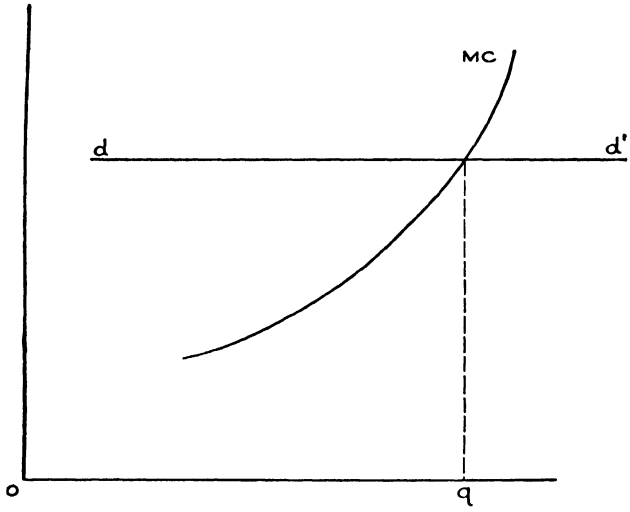

Figure 45

for the firm at any output will simply equal the sum of factor payments made to secure that output, divided by the output; the marginal cost of production will be the addition to factor payments for a unit addition to output.

This marginal cost will tend to increase beyond a certain output if "the firm" is in effect an imperfectly divisible unit against which the proportions of hired factors vary. That is, average and marginal costs of output will rise as output is extended beyond a certain point (even in the long run), because of the declining efficiency of management beyond a certain scale of firm. Rigidly mixed doses of the three factors (*A*, *B*, and *C* in the 5-3-1 proportion) will produce successively less output as diminishing returns are encountered. With this progressive rise in marginal cost, the competitive firm will set output at the point where marginal cost is equal to the price of its product, as in Figure 45. At this output of goods, it will employ a *corresponding* amount of factors, necessarily in fixed ratio; for example, it may employ 500 units of *A*, 300 of *B*, and 100 of *C*. By determining its output, it thus also exactly determines its demand for each factor, given the price of its output and the set of prices for the three factors. The *relative* quantities of the various factors employed is set by

the fixed technical ratios which rule. *The absolute quantities employed depend upon the relation of the total money price of a composite dose of factors to the money price of the good produced, upon the productivity per dose of factors, and upon the rate at which this diminishes with increasing output.* The individual firm determines only the quantity of factors to employ, and cannot affect their market prices. It will produce only if price covers average costs, but if we view one firm alone, an excess of price over average costs is quite possible.

We may now drop the somewhat artificial assumption of fixed technical ratios among factors, and consider a second situation, where it is recognized that the firm may vary the proportions in which factors are employed in order to obtain any chosen output. In actuality, various factors are substitutes for each other; labor, for example, may be substituted for capital, or capital for labor. Instead of using factors *A*, *B*, and *C* always in a 5-3-1 ratio, the firm may experiment with other ratios, such as 6-2-2, 4-4-1, and so forth. At each level of output, it is free to choose a specific proportion among the factors and will presumably make its choice in such a way as to minimize the cost of the output. The proportions in which the factors are employed will thus obviously depend upon their relative prices.

The choice is also necessarily conditioned, however, by the nature of the substitutability of one factor for another. Such substitution relationships observe certain regularities under substantially all conditions of production. In general, the proportions of any two factors may be varied (increasing the amount employed of one, decreasing that of the other) to produce a given constant output. But starting from any initial combination, successive *unit decreases* in the amount of any one factor will require successively *larger increases* in that of the other in order to maintain output. This fact is sometimes expressed by saying that the *marginal rate of substitution* of any factor *A* for any other factor *B* must increase as *A* is progressively substituted for *B*. Such a typical relationship is illustrated in the following table, which shows the various alternative combinations of two factors, *A* and *B*, that may be employed to obtain a given output of 1000 units of a certain product:

COMBINATIONS OF FACTORS *A* AND *B* FOR ($Q = 1000$)

Units of factor *A*	Units of factor *B*
40	61
45	55
50	50
55	46
60	43
65	41

This table says that the firm can produce 1000 units of output by employing 40 of *A* and 61 of *B*, *or* by employing 45 of *A* and 55 of *B*, and so forth. The firm can use less of *A* and more of *B*, or vice versa. If either factor is successively decreased in quantity, however, the other factor must be increased at an increasing rate. Thus a decline in the use of *A* by 5 from 65 to 60 can be compensated by an increase in *B* of 2, but a further decline in *A* of 5, from 60 to 55, requires an increase of 3 in *B*. (The marginal rate of substitution of *B* for *A* rises from 2-5 and 3-5.) In a sense, each productive factor tends to encounter diminishing marginal returns or productivity as it increased against any other. This tendency is, of course, an effective check against a firm using any one factor entirely and to the exclusion of others.

The relation shown in the preceding table is charted in Figure 46. The curve labeled "$Q = 1000$" shows the various proportions of *A* and *B* which may be employed to produce 1000 units of output. A similar curve, of similar shape, may be drawn for every other alternative level of output—thus the curves labeled "$Q = 950$" and "$Q = 1050$." Each such curve is an "isoquant," or curve showing combined factors quantities necessary to produce a constant quantity of output.[3] With factor *B* on the vertical axis, its slope, which increases as *B* is increased, measures the marginal rate of substitution of *B* for *A*. At any level of output, the general principle of the increasing marginal rate of

[3] We neglect here, and thus leave for more advanced treatments, the effect of the indivisibilities of hired factors (encountered short of optimum-scale outputs) and of the existence of diminishing returns of the composite of hired factors against a fixed "firm" (if any) on the substitution conditions among factors. See Boulding, *op. cit.*, Chap. 23.

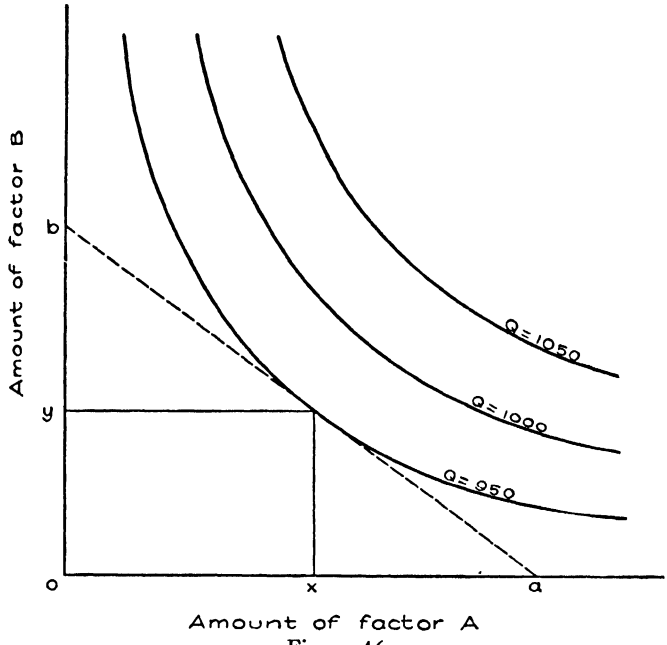

Figure 46

substitution of one factor for another holds, and it is quite as applicable to the covariation of three or more as to that of two factors.

What does this matter to the firm buying factors? Principally this—as the firm adjusts its output so that marginal cost will equal the price of its output, it must also choose that proportion among the factors which keeps the cost of every output at a minimum. And this minimization of cost—a choosing of the *cheapest* combination of factors at each output—is reached by balancing the relative prices of various factors against their marginal rates of substitution one for another. *To minimize costs, the firm will necessarily choose such a combination of any pair of factors that the marginal rate of substitution of A for B is equal to the ratio of the price of B to that of A.*

This principle may be illustrated simply from the preceding table, which gives certain substitution ratios for increments in either of two factors. Suppose that *A* costs $3 per unit and *B*

costs \$5 per unit. With 50 units of *A* and 50 of *B*, the total cost of 1000 units of product would be (\$150 + \$250) = \$400. Consider now substituting one factor for another in either direction. The firm will not substitute 5 units of *B* for 5 of *A* (moving upward in table) since this would increase costs by \$10. But it will substitute 5 units of *A* for 4 of *B*, thus reducing costs by \$5, and it is willing to substitute 5 *more* of *A* for 3 of *B*, moving to the combination 60-43, and leaving costs unchanged at \$395. It will *not* substitute another 5 of *A* for 2 of *B* (moving to the 65-41 combination), however, since costs would rise to \$400 again. The firm therefore finds the optimum proportion at 55-46 or 60-43, or in the range between these points, where the *marginal rate of substitution* of *A* for *B* is 5 to 3. It will be noted that this rate is the inverse of the *A-B* price ratio of 3:5. At *every* output, the firm will make this sort of adjustment in the proportion of factors employed. It will minimize the cost of securing any output by employing factors in such proportions that the marginal rate of substitution of any factor *A* for any other factor *B* is equated to the ratio of the price of *B* to the price of *A*.

This solution may be represented diagrammatically in Figure 46, where each of several isoquants show the varying factor combinations required to produce given outputs. The varying slope of any isoquant shows the varying marginal rate of substitution of *B* for *A* as the proportions are varied. The dotted line *ab* is drawn to represent the inverse of the *B*:*A* price ratio—that is, the ratio of the price of *A* to that of *B*, when both of their prices are given to the firm. (The distance *ob* represents the amount of *B* which can be bought with a given amount of money; the distance *oa* the amount of *A* which the same amount of money will purchase. Then the line *ab* represents all combinations of *A* and *B* which can be bought with the given amount of money, and its slope represents the ratio of the price of *A* to that of *B*.) Where such a line *ab* is tangent to a given isoquant, then the marginal rate of substitution of *B* for *A* is equal to the inverse of their price ratio, and at this point the firm finds the minimum-cost combination of factors for producing that output—thus in Figure 46, *ox* of *A* and *oy* of *B* to produce 950 units. When the isoquant is represented as a continuous curve, the best com-

bination is a point on this curve. When it is represented as a series of discrete points, as in the preceding table, the cost-minimizing combination appears to lie in a range between two points.[4]

The firm may thus be regarded as *simultaneously* adjusting its output so as to equate marginal cost to price, and the amounts and proportions of factors hired in such a way as to keep costs at a minimum. The unit price of each factor is still given to the firm. The amount it purchases of each factor depends both upon the relation of the price of its output to the general level of factor prices, and upon the relations among the prices of various factors. In effect, the firm's action is generally the same as that suggested for the case of fixed proportions of factors but involves an added decision on the proportions of various factors to employ. The firm decides to purchase a determinate quantity of each factor after balancing a complex of considerations including the prices of all factors, the price of its output, the substitution relations among factors, and the variation of productivity of factors with variation in the firm's rate of output.

In this circumstance, is it meaningful to speak of a firm's "demand curve" for any one factor, which would show the amount of the factor it would purchase at each of a range of prices? Assuming to be given the price for the firm's output and the prices of all other factors, we can trace the effect of the changes in the price of some one factor upon the quantity of it which a firm will buy. Thus, under the assumptions noted, we might plot a firm's demand curve for labor. As the price of labor is reduced, other factor prices remaining constant, the firm should simultaneously (1) substitute labor for other factors, to the end of minimizing cost of output, and (2) vary (*i.e.*, extend) output to *keep* marginal cost equal to price. Thus a reduction in the price of labor should increase the quantity a firm will purchase. The line showing the resulting relation of quantity purchased by a firm to price could be labeled the firm's demand curve for labor. Thus there might be a firm's "demand curve" for labor, for capital, or for land. In a sense each firm,

[4] See J. R. Hicks, *Value and Capital*, Chaps. 6 and 7; and Boulding, *op. cit.*, Chap. 23, for a more detailed treatment.

in maximizing profits, is continually hiring each factor at the point where this "demand curve" intersects the supply curve for the factor.

PURCHASE OF PRODUCTIVE FACTORS BY THE INDUSTRY

Let us now shift from a single firm in pure competition to an industry of such firms and inquire how the combination of their actions affects the determination of factor prices and employment. We have indicated that where there are uniformly competitive conditions in factor markets the supply of each factor to the firm will be perfectly elastic. Each factor price is given regardless of the amount taken, because any firm is too small a buyer to affect the price of any factor. If we assume that there is one homogeneous, economy-wide market for each factor, the same should be true of the small industry. That is, the total purchases of any factor by the industry would be so small that the industry cannot influence its price—the price of each factor is given to the industry. This may at any rate be assumed for purposes of illustrative argument.

The only new matter to be discussed for the industry, therefore, is the interaction of competing firms' demands for factors. This interaction is linked with the familiar process of determination of long-run price and output for a competitive industry. First, the process by which any *given* number of firms select outputs for which marginal cost is equal to price results in the determination of an industry price for which this is simultaneously possible for all firms. At such a provisional equilibrium point, all firms would be employing the various factors in determinate amounts and proportions. Second, the existence at this point of any excess profit or net loss induces entry or exit, until each firm is driven to the point of producing at minimum average cost and to selling at a price equal to this cost. As this long-run equilibrium point was reached, each firm would correspondingly adjust its purchases of each factor according to the principles noted, and a given aggregate industry purchase of each factor would result. In this equilibrium the price of the industry's output under the pressure of free entry equals minimum average cost; there are no true profits; and the total payments to factors

necessarily equal the total sales receipts of the industry. For an industry of competitive firms purchasing factors, the pressure of competitive price-output adjustments tends to force the industry purchases of any factor (given its price) to such a determinate long-run point.[5]

Can we now meaningfully speak of an industry's long-run "demand curve" for any factor—*e.g.*, labor. If we assume the prices of other factors to be given, we can trace the effect of a change in the price of one factor on the amount of it the industry buys. In this case, a reduction in the price of a factor will (1) induce its substitution for other factors to some extent, and (2) lower the minimum average unit costs of production of every firm, allowing (with a given demand for the industry's output) a larger total output. The addition to output will be supplied by the entry of new firms. The combined adjustments of substitution and entry should result in a determinate extension of the purchases of the factor by the industry. The response of demand for the factor to change in its price will evidently depend strongly upon (1) the elasticity of demand for the product of the industry, (2) the degree or rate of substitutability of the factor in question for others, and (3) the conditions of entry to the industry—whether added firms will be as efficient, or less or more so, than those preceding them. Reflecting all these conditions, a line can be drawn showing the relation of an industry's purchases of a factor to its price, and this might be called an industry's demand curve for a factor—*e.g.*, labor. It should be strongly emphasized, however, that such a curve is *not* simply an addition of individual firms' demand curves for a factor, since several industry-wide adjustments not taken into account for the firm's curve do enter into the definition of an industry demand curve.

[5] The number of firms in such an equilibrium is determinate and finite provided the firm reaches minimum average costs at some finite output, on either side of which its average costs would be higher. The firm's U-shaped long-run average cost curve in turn results from the fact that the "firm" is not perfectly divisible or variable in size, and possibly from the indivisibilities of hired factors at very small firm outputs. We have neglected here, however, the effect of both sorts of indivisibility on substitution conditions. Recognition of these effects will not significantly modify our conclusions concerning industry equilibria.

PURCHASES OF PRODUCTIVE FACTORS BY A COMPETITIVE ECONOMY

In investigating the industry in pure competition as a purchaser of factors, we have been able to elicit certain principles governing the quantity of factors such an industry will take at given prices. But because the prices of factors may be assumed given to any one industry, we have not yet come to grips with the determination of factor prices, and with numerous important related matters. We may now carry our analysis a step further by considering the purchase of factors *by a competitive economy—i.e.*, by an economy made up of a large number of industries in pure competition. We thus broaden our viewpoint to consider the interactions of the decisions of all buyers of all factors in such a competitive situation. Although the assumption of purely competitive conditions in all pricing is unrealistic, the simultaneous consideration of the whole economy brings us much closer to real problems than any analysis of individual industries ever could. The pricing of any productive factor is essentially an economy-wide phenomenon and cannot be fully understood by a particular analysis of individual industries. By considering the whole economy, we bring into focus two essential considerations which must condition the outcome of income distribution and employment: (1) the supplies of the various factors cannot be regarded as perfectly elastic to the economy—all factors are scarce and relatively inelastic in supply, and their prices thus remain to be determined; (2) the payments made to factors for the economy as a whole are the incomes from which demands for goods are forthcoming—the relations of total factor payments to total demand for goods must thus be made a part of the explanation of factor pricing and employment. Let us consider these problems—*i.e.*, factor-price determination and general equilibrium for the economy—in turn.

Supposing an economy of purely competitive industries buys all factors in purely competitive factor markets, let us first also suppose, somewhat arbitrarily, that there is a given constant flow of money purchasing power or aggregate demand price for all outputs, regardless of the level of money payments to factors.

An economy of industries with a given interrelated family of money demands for their products (stemming from a constant aggregate flow of purchasing power) must thus buy a group of factors each of which is in relatively inelastic supply. To simplify the problem further at first, let us suppose that for the economy each factor is in perfectly inelastic supply. That is, each factor is available in a fixed quantity per unit of time, and all of it can be bought at any price the market offers.

In these circumstances, the money price of each factor will necessarily move to such a level that the total demand for it by all industries in the economy will just equal the fixed supply—*i.e.*, to the level consistent with full employment. More generally, the family of money factor prices will arrive at such mutually consistent levels that this will be true simultaneously for all of them.

Let us trace the process which is implied. To any given *arbitrary* set of money prices for factors, all industries would tend to adapt themselves competitively, until each industry was in long-run equilibrium (the interrelated prices for various industry products having reached a stable mutual adjustment). At this level there would tend to be a determinate aggregate economy-wide demand for each factor. But at such arbitrary money prices, this aggregate demand for any factor might either exceed or fall short of the fixed supply. If the demand exceeded the supply, the money factor price would tend to rise until, with a tendency toward substitution against the factor and restriction of output, demand was equated with supply; if demand fell short of supply, competition among sellers in the factor market would drive the money factor price down, until with substitution in favor of the factor and extension of output, demand was equated with supply.

Such an adjustment would proceed simultaneously for all factors until a mutually consistent set of money prices for them was established. For the whole economy there would be a covariation of factor prices, factor proportions, commodity outputs, and commodity prices until a position was reached where there was full employment of every factor and where simultaneously every industry was in long-run equilibrium with respect to cost-price relations and factor proportions. In this case

(which, it must be emphasized, supposes universal pure competition, a given *constant* rate of flow of money purchasing power, and perfectly inelastic supplies of all factors) the end result would be full employment of all factors, price–average-cost equality for all industries and firms, and a total of money payments to factors which was just equal to the total flow of money sales receipts to firms. No true profits (in excess of total wages, rents, and interest) would be earned.

Two aspects of the resulting equilibrium of prices for the economy may be emphasized. First, the economy reaches an equilibrium *money factor price level*—or average money price of factors—such that with the given flow of money payments all factors are employed. This equilibrium level of average money prices will depend upon the relation between size of the flow of money payments and the size of the aggregate supply of factors. Second, there is struck an equilibrium of *relative* factor prices—of the ratios among factor prices—such that, with uniformly full employment, all firms may employ factors with marginal rates of substitution which are in balance with price ratios. This means that for the economy as a whole the relative prices of any pair of factors necessarily assumes a ratio the inverse of which is equal to the marginal rate of substitution of one for the other *for the whole economy*. It follows that, given certain techniques and corresponding substitution relationships, the relative price of any factor tends to become lower as that factor is more plentiful relative to others. As its quantity is increased, its marginal rate of substitution for other factors throughout the economy increases, and the ratio of its price to other factor prices must fall. The competitive wage of labor, for example, tends to fall relative to rents as the labor force working with a given supply of land is increased.

An economy-wide "demand curve" for any factor *could* be constructed by supposing arbitrary variations in the fixed supply of that factor, other factor supplies remaining unchanged. The response of its price to such variations would reflect the economy-wide composite of adjustments to the change in supply, as the economy moved from some initial general equilibrium position to a second.

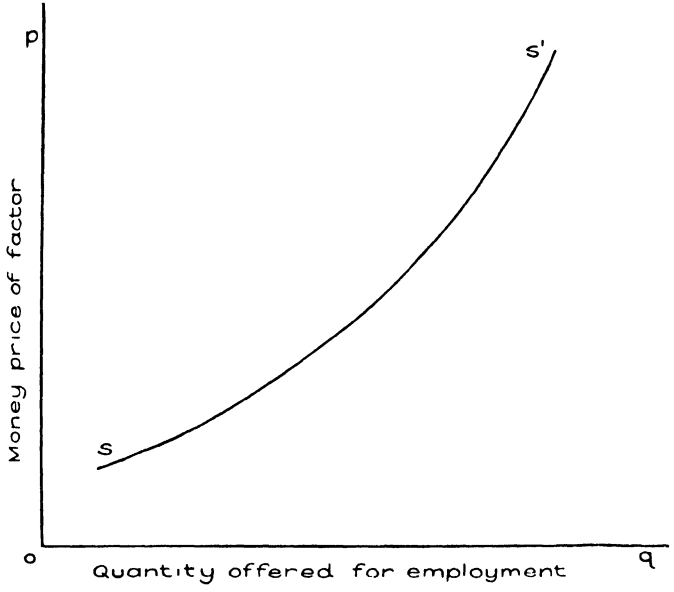

Figure 47

So far we have proceeded on the assumption that the supply of each factor is in perfectly inelastic to the economy. Suppose, instead, that each factor has an elastic and sloping supply curve, indicating that various amounts will be offered at various money prices, as in Figure 47. This money-price supply curve may be viewed provisionally as independent of the prices of commodities (*i.e.*, given regardless of the buying power of the dollar), while we retain the assumption of a given aggregate money demand for aggregate output. (It is doubtful that stipulations for money wages, rents, etc., *are* independent of commodity prices, but we may investigate the implications of this condition as a step in analysis.) Under the assumed conditions, the burden of the preceding analysis applies without serious revision, but with one substantial addition. In bidding for factors now, the economy of firms would determine not only the money prices of all factors but also the amount of each which would be employed. With a given constant flow of money purchasing power demanding the outputs of firms, a consistent set of prices for all factors

would be arrived at by bidding, and these prices would simultaneously fix corresponding rates of employment. These rates of employment would necessarily depend upon the relation of the aggregate flow of money purchasing power to the money factor prices, stipulated in the factor supply curves, which were required to induce various amounts of the various factors to work.

Should the money supply prices of factors, as shown in their supply curves, in general be high relative to the assumed given flow of purchasing power, substantial unemployment could result. The specific level of employment of any factor (relative to others) would similarly depend upon the relative height of its money supply price. This much, however, is evident. If some factors remained unemployed in this situation, the unemployment would be "voluntary," in the sense that it would result from the unwillingness of unemployed factors to accept as low a money price as a competitive market would offer. An equilibrium with some amount of unemployment would be quite tenable, provided there were a given flow of purchasing power and inflexibly stipulated money factor prices. At this equilibrium, as before, the total of factor payments would equal the total receipts of enterprise, with no true profits for competitive industry. But net unemployment could result and if it did, it would be the result of arbitrary money factor prices, here arbitrarily assumed to exist.

Two premises, however, should be re-examined: (1) that the aggregate money demand for goods is given independently of money factor payments, and (2) that factors stipulate for given *money* prices regardless of the level of commodity prices, thus giving rise to independent money-price supply curves.

THE MUTUAL DETERMINATION OF AGGREGATE SUPPLY PRICE AND AGGREGATE DEMAND PRICE

The aggregate money demand price for all commodities, or aggregate of receipts of all enterprises, is of course not given independently. It is related to and derived from the aggregate supply price of commodities, or aggregate of money income payments to factors, including any residual profits. Each feeds the other in a perpetual circular flow through time. There is a

time sequence involving an initial aggregate money demand for commodities, a derived money income payment to factors, a further derived demand for commodities, and so forth indefinitely. Thus let D_1 represent aggregate money expenditure on commodities in any initial period, and S_2 the money payments to factors derived from it. Let D_3 be the money demand derived from S_2; S_4 the income derived from D_3; D_5 the demand derived from S_4; etc. Then D_1 influences S_2, which influences D_3, which influences S_4, which influences D_5, which influences S_6, etc. One analytical task is to explain the character of this endless-chain relationship—that is, to explain the behavior of the flow of money payments over time.

Without explaining the relation, however, it is evident that various sorts of relationship are conceivably possible. One is that the flow of money payments proceeds at a *constant* rate—that is, that enterprises always make a total of money income payments to factors (including profits) just equal to their total money receipts (aggregate demand price), and that income-recipient factors and their owners (including profit-receivers) in turn always offer a money demand for goods equal to what they have received.[6] There would thus be a stable and self-reproducing circular flow of payments; in effect, $D_1 = S_2 = D_3 = S_4 = D_5 = S_6 \cdots = D_n = S_n$. Supply would regularly create its own demand. Another possibility is that income recipients regularly spend less money than they receive. Then $D_1 > D_3 > D_5 > D_7$, and so forth, and the flow of money payments dwindles steadily through time. Or it is possible that $D_1 < D_3 < D_5 < D_7$, etc., so that the flow of money payments continually increases over time. Or a fluctuating behavior of D through time is quite conceivable.

Let us center attention first on the possible case where supply does exactly create its own demand—where enterprises always pay out in money incomes an amount exactly equal to total money receipts, and where income recipients always spend on the output of enterprises an amount exactly equal to what they have received. In effect, enterprises conduct no net *hoarding* of

[6] And this in spite of any adjustments of hired-factor prices which may occur.

money, and neither do income recipients (they do not add to their unspent money balances through time), and similarly, neither enterprises nor income recipients *dishoard* any money, or otherwise supplement the flow of money payments from other sources of funds. *If* this is the case, then $D_1 = S_2 = D_3 = S_4 = D_5$, etc. Or, perhaps more importantly from the standpoint of enterprises, $D_1 = D_3 = D_5 = D_7 \cdots = D_n$. That is, the flow of money demands for commodities would remain constant at any initial level. Where this is true, then, *the situation is just the same as if aggregate demand were independently given at an arbitrary level.* The circular interdependence of incomes and demands actually makes any level of money demands permanent.

So far as this condition is found, the arguments of the preceding section, developed on the supposition of an arbitrarily given flow of money demand payments, may be applied here without modification. The competitive economy, that is, would adapt itself to the given constant flow of money demand, and after a sufficient passage of time would reach long-run equilibrium corresponding thereto. The general properties of this equilibrium would be that for all industries price was equal to average and marginal cost of production, and that the total of factor payments (excluding any true profits) was thus just equal to the total receipts of all enterprises. The level of money factor prices would thus be brought into a balanced relation to the general level of commodity prices. The size of employment and of aggregate output, and the absolute level of factor and commodity prices, would depend upon the size of the money flow, *and upon whether factors were held for arbitrary money prices.*

If all factors were offered in perfectly inelastic supply without withholding for any specific money price, there would automatically tend to be full employment of all factors and a corresponding level of output. Money prices of factors and commodities would simple move to such a general level as to make this possible with the given flow of money payments. If, on the other hand, factors generally were held for given money prices, and these were above the levels necessary for full employment with the given money flow, there would be perpetuable unemployment and restriction of output. This would be attributable

directly to arbitrary money prices for factors which were "too high" for the size of the money income flow. In either case, however, supply would create its own demand in the equilibrium situation. Total expenditure on commodities would always equal the total outpayments made in producing them. And, since all goods come to be produced in long-run equilibrium without excess (or true) profit, total revenues would thus just cover total costs, and every enterprise could always just "break even" in selling its equilibrium output, thus having no desire to expand or retract.

The idea of equilibrium unemployment in this situation, moreover, seems largely arbitrary or accidental. With a given money flow disposed to be self-perpetuating regardless of factor-price adjustments, unemployment appears as the incident of an arbitrary stipulation by a certain factor or factors for a money price above the level where full employment can be had. Such arbitrary stipulations may occur in the actual world. But in the situation we are discussing, this unemployment would be easily eliminable. On the one hand, it could obviously be remedied by increasing the money flow in the economy—as for example by governmental creation and expenditure of new money. This would overcome any unemployment caused by simple arbitrary rigidity of the money prices of factors, provided that such money prices were not adjusted upward as the money flow increased.

More generally, such unemployment would not persist if the factors or their owners ceased holding out for arbitrary money prices, and adjusted their stipulations with an eye to the *real income,* in terms of commodities, which their money incomes would buy. If this were the case, the economy-wide "supply curves" for factors of production would intrinsically be expressed in "real" terms—that is, each one would express the relation between the amount of real goods implicitly offered in payment to a factor and the quantity of the factor which would be offered for employment. Actually, of course, enterprises would pay money prices and factors would hold out for money prices. But if the decisions of factors or their owners turned on "real income" considerations, we would find that the money-price supply curves of factors were not arbitrarily given, but shifted progres-

sively in response to shifting commodity prices until a final equilibrium was reached.

In this equilibrium it would be true not only that the money price of each factor had reached a level where all units of the factor offered for employment at this money price were employed, but also that all units offered for the real income which this money price could purchase were employed. There would still presumably be unemployment, but it would be truly "voluntary," in the sense that all unemployed units would refuse to work for the real income which the market in equilibrium would pay. Such a complete elimination of "involuntary" unemployment, with progressive adaptation of the money prices of factors to their "real" withholding prices, is quite possible so long as the flow of money payments is unaffected by the process of factor-price adaptation. And this is evidently so under the condition assumed: that the total of money income payments by enterprises (including profits in transitory periods of disequilibrium) are always fully respent and become effective money demands for goods.

If this were true, and if in addition all suppliers of factors thought ultimately in "real" terms, considering their prices in relation to commodity prices, the specific level of the flow of money payments would be of no consequence. The same real equilibrium of employment, output, and *relative* prices of commodities and of factors would be struck at any level of money income. The only thing influenced by the size of the money flow would be the absolute level of money prices for commodities and factors.

A world where the flow of money payments was in effect self-perpetuating would obviously be a very simple economic world. The level of employment of all factors, the level of output, the relative prices of commodities, and the relative shares of total income received by various factors could work themselves out to a long-run equilibrium without setting up any cumulative or frustrating disturbances in the flow of money payments. The flow of payments could be a neutral consideration in the economy, and it would not be surprising if people came to calculate alternatives primarily in real terms, viewing money strictly as a "veil."

Given this view, a remarkable sort of balance would tend to be struck by a purely competitive economy. Perhaps the most significant aspect of this equilibrium is that there would be for each factor, regardless of the size of money payments, a determinate level of employment involving no involuntary unemployment. The relative shares of income going to the various factors would depend upon their relative productivity under prevailing techniques and upon their relative "scarcity."[7] Otherwise we cannot say *a priori* how large the wage bill, for example, would be relative to the total of interest or rent payments. But we can say that in the situation noted (self-perpetuating flow of payments, universal pure competition, and calculation of economic alternatives primarily in real terms) the result would be quite automatic, free from any deliberate control by individuals, and definitely determined at a unique competitive equilibrium level.

There is a certain attractiveness to such concise simplicity, and this appeal may explain the tendency of some economists to confuse this simple model world with the one we actually have. Yet it is evident that the actual economic world departs from this model in all three essential respects. That is, (1) the commodity and factor markets in the actual world are not purely competitive, (2) people do not entirely disregard money prices or "see through" them to their real counterparts, and (3) it is not true that supply always just creates its own demand and thus insures a self-perpetuating flow of payments.

On the last point, an obvious alternative is that supply does not create an equivalent amount of demand in every circumstance; that after a given payment of money incomes by enterprises, a smaller or larger amount is spent on the goods produced. There may be, that is, *net dishoarding* or *net hoarding* on the part of income recipients, thus causing the flow of payments to wax, to wane, or to fluctuate. In this event, the circular interdependence of factor incomes and commodity demands creates genuine problems. Without entering in this chapter into an analysis of why and under what conditions the flow of money payments will remain constant or change, we may comment on

[7] Due to natural limitations (as in the case of resources) or to psychological limits on the inclination to work (as in the case of labor).

three possible relations of aggregate money income payments to aggregate money demands for goods:

1. The aggregate money demand for goods continually tends to rise above the aggregate of income payments, so that the flow of money payments rises progressively; that is, $D_1 < D_3 < D_5 < D_7$, and so forth, because of dishoarding or of other supplementation of the flow of money payments.

   Such a condition seems to characterize periods of cyclical expansion (from depression to prosperity) but seems unlikely to persist indefinitely.
2. The aggregate money demand for goods tends continually to fall below the aggregate of money income payments, so that money demand falls progressively. That is, $D_1 > D_3 > D_5 > D_7$, and so forth, because of net hoarding or other subtraction from the flow of money payments.

   Such a condition seems to characterize periods of cyclical contraction (from prosperity to depression), but also seems unlikely to persist indefinitely.
3. The aggregate money demand for goods will always be less than the aggregate of money income payments *for high levels of employment, but will become equal to income payments at some lower level of employment of factors*. Thus, for example, between "100 percent" employment and "80 percent" employment, $D_1 > D_3 > D_5 > D_7$, etc.; but at "80 percent" employment $D_1 = D_3 = D_5 = D_7$, etc.; and at *below* 80 percent employment, $D_1 < D_3 < D_5 < D_7$, etc.

   This is a possibility often emphasized as a probability in recent theoretical discussion. It is suggested, that is, that in given circumstances there may be an "equilibrium" level of employment at which the flow of money payments will remain constant but that at higher or lower levels of employment the flow will tend to decrease or increase. Such a condition is sometimes held to represent a long-run tendency, aside from temporary cyclical fluctuations in the flow of money income.

What are the implications of these three alternative situations for the determination of levels of employment and of income payments to factors?

If aggregate money demand for goods continually exceeds the money income payments from which it is derived, and monetary expansion thus persists, a continued upward adjustment of commodity prices, factor prices, and (to a limit) output and employment is indicated. The demand curves for the outputs of firms will continually shift upward and to the right, causing the money prices of output to rise and inducing firms to attempt to extend output. In so doing, firms must bid for more factors, forcing money factor prices up correspondingly. Employment (and hence total output) *may* increase so long as there are unemployed factors, and so long as money factor prices do not rise as rapidly as the money demand for their services increases. But if the money prices which unemployed factors require to employ them rise with the rising money flow simultaneously and at a sufficient rate, employment will not increase (and might even decline).

In this situation of expanding money payments, however, there is a strong probability that "full employment" (*i.e.*, no involuntary unemployment) will after a number of periods for adjustment be reached and maintained.[8] Alternative possibilities of "overemployment" (factors working for payments which turn out to have a lower real purchasing power than they wish to accept) and "underemployment" may, however, result either from arbitrary money price stipulations by factors, or from miscalculations by factors of the movement of the money price level for commodities.

One specific alteration in the over-all distribution of income may in any event result. As commodity prices are bid up by the rising flow of money payments, factor prices *in competitive factor markets* may tend to rise *later*—that is, their rise may continually tend to lag behind that of commodity prices through time. This would tend to be the case if goods were sold only

[8] This is because, with the necessity of progressive upward shifts in the money-price supply curves of factors, there is every opportunity for this shifting to lag behind the rise in money income until the employment of all factors which desire employment at going levels of "real" price is accomplished.

after passage of a discrete time interval after production (and thus at higher levels of demand and price than ruled when the goods were produced) and if payments made to factors by firms were always calculated on the basis of demands ruling at the time of production. Where such a lag exists, enterprises will continually tend to realize a *profit*, in the form of a windfall from rising prices, over and above the return to their labor and capital, and thus, in spite of pure competition, to receive an excess-profit share of income at the expense of hired factors. Continued monetary expansion and price inflation may thus tend to distort the distribution of income in favor of an extra profit share for enterprises.

Let us consider now the reverse case, where aggregate money demand continually falls below money income payments to factors, so that the flow of money payments through the economy steadily dwindles. Here enterprises find the money demands for their products continually shifting downward and to the left, and the prices of their outputs falling. They thus individually tend to (1) restrict outputs, and (2) employ less factors. What will happen to the level of employment and output? This will depend primarily on the rate of downward adjustment of money factor prices (and enterprise costs) which is induced by the diminution of enterprises' money demands for them. Should money factor prices remain rigid (or relatively so) because of unwillingness of factors or their owners to accept smaller money payments, a progressive accumulation of unemployment seems inevitable. For full employment and output to be maintained, it would be necessary for money factor prices to fall freely and fluidly with the fall in money payments, and to continue to do so as long as the flow of such payments continued to decline. If the money price stipulations of factors are "sticky" or change only with a lag in response to declining money payments, a sort of "frictional" unemployment will tend to result. Also, following the argument of the preceding paragraph, a lagging adjustment of factor to commodity prices will tend to cause chronic *net losses* (negative profits) for enterprises. Persistence of such declining money income may thus tend to cause business failures and a breakdown of the enterprise system itself.

Let us now turn to the third case, where it is supposed that

the flow of money payments tends to become stable only at a certain (provisionally less than full) level of employment, declines if employment is greater, and rises if employment is below this level. The result in this case would be quite simple if the money prices of factors were arbitrarily set at certain levels, as they might be, for example, because of "traditional" notions of satisfactory money payments. Suppose that "80-percent employment"[9] were the level where aggregate money demands would equal aggregate money income payments to factors. Then money payments would simply move to the level necessary to employ 80 percent of factors at the given money prices. If employment were higher, the money flow would diminish; if it were lower, the money flow would increase, reaching equilibrium where 80-percent employment was obtained. This would suppose, however, arbitrary money factor prices set without regard to the real income earned thereby.

What would ensue should the suppliers of factors attempt to adjust their money prices downward to overcome the unemployment—if unemployed factors were willing to work for smaller real incomes than those earned at the 80-percent-employment equilibrium? In this event, money factor prices would tend to fall, but to no avail. Because so far as enterprises therefore attempted to employ more factors, the flow of payments would tend to decline, making such additional employment unprofitable. Where monetary equilibrium is virtually obtainable only at less than full employment, therefore, and where money factor prices are *not* rigid, there would tend to be a continual decline in the factor and commodity price level, with a continuing margin of involuntary unemployment. But rigidity of money factor prices can "save" this situation and allow an arbitrary monetary equilibrium with a given quantity of unemployment.

SUMMARY

In this chapter we have set out to analyze the distribution of income in an enterprise economy. Narrowly, this analysis at first

[9] *I.e.*, 80 percent, on the average, of the factors which would wish to be employed at the real prices payable to them in a competitive general equilibrium.

seems to be concerned simply with the determination of the prices of various factors of production. Broadly, it turns out to be concerned also with the determination of the level of employment of all factors, of the shares of income received by each, of the size of the flow of money payments through the economy, of the level of output, and of the general level of money prices of factors and commodities. These matters generally have been investigated for the very simplified situation of a purely competitive economy, where all markets, for both commodities and factors, have many buyers and many sellers. Subject to such drastic simplification, certain general tendencies may be analytically isolated and observed. What, in effect, have we learned from this analysis, and how much of it is applicable to the real (and other than purely competitive) economy?

Before summarizing the analysis, it may be useful to review the assumptions upon which it rests—assumptions describing the structure of all markets in the economy. A first assumption is that there is a purely competitive market for each and every commodity, in which many sellers of an identical product sell to many buyers. A second is that there is a single, large, purely competitive market for each of several factors of production. Within each factor market all enterprises purchasing the factor are competing buyers, and all units of the factor are competitively offered for sale. There is no differentiation among various units of any factor. There are thus just three factor markets in the whole economy, and each is purely competitive. Within the market for any factor, we abstract from the effects of subdivisions by regions or on the basis of type or quality.

In this simplified setting, certain tendencies stand out clearly. First, beginning with any going level of money demands for their products, an economy of competitive firms will bid for the services of various factors, at whatever their current money prices may be, purchasing them in *proportions* and *amounts* such that (1) the marginal rate of substitution between any pair of factors will be equal to the inverse of their price ratios, and (2) that the marginal cost of producing output will be equal to the price of output. Enterprises will continually adjust the proportions and amounts of factors employed to maintain these basic equalities as the money prices of factors and commodities

shift about. And these money prices will in fact tend to shift toward some sort of equilibrium. That is, commodity prices in all industries will tend to be adjusted, by pressure of entry or exit, toward equality with minimum average costs of production, with an accompanying adaptation of the employment of factors. Also, equilibrium money prices for all factors will tend to be struck, balancing the total demands for factors against the amounts of money which factors or their owners require in order to supply them for use. (Such equilibrium money prices may also—although not necessarily—involve the balancing of factors' supply prices in terms of real good against the real incomes implicit in their prices.) As factor prices are established, enterprises adapt their amounts and proportions of employment to these prices. These money factor prices become the costs, of course, to which commodity prices are successively equated.

Now in turn, the income payments made to factors "circulate" through the economy and determine the money demands for the firms' products. If the relation of such demands to incomes is other than that of identity, it is evident that any tentative equilibrium at which the economy of enterprises and factors may arrive is continually upset by a changing money flow and changing money prices. If, however, it happens that all money income payments to factors are exactly respent, the money demand for commodities remains stable through time and the economy can become fully adapted to a given flow of demand. In this event, all the adjustments outlined in the preceding paragraph could proceed to a point of general equilibrium for the whole economy. There, all commodity prices would equal marginal and minimum average costs of production. There would be no true profits—all income would go to hired factors of production. This equilibrium could be reached with such a relation of the general level of money prices for factors and commodities to the flow of payments that there was full employment—that, at the "real wage" paid each factor, all units of that factor which were offered for work were employed. The general level of money prices should in fact move to such a point provided factors do not make arbitrary demands for certain money payments and thus "peg" the money price level. (But such arbitrary stipulations could evidently limit employment in situations of

a given or inflexible flow of payments.) In the full general equilibrium of the economy with full employment, the absolute money prices and incomes of factors would be of no consequence. Their *relative* prices and the relative shares of income they received would depend upon their relative scarcity, their "supply prices" as expressed in real terms, and their relative usefulness in production.[10]

If, on the other hand, a stability of the money flow through time is not automatically obtained, such a general equilibrium may not be found. Departures may involve not only rising or falling general money price levels, but also departures from full employment and distortions of income distribution through windfall profits or losses to enterprises. In such situations, moreover, adherence to arbitrary money price stipulations by factors becomes increasingly probable, and may be instrumental in stabilizing the economy, possibly with chronic unemployment.

This is a review of the general function of the distribution and flow of income in a capitalist economy, developed under the drastic simplifying assumption of universal pure competition. Certain general tendencies seem to stand out clearly. We may now inquire in what degree our conclusions must be modified in order to apply more readily to the real world. A number of additional steps in analysis are evidently required to adapt the preceding analysis to the more complex situations found in the real economy, and also to extend the analysis to take in further ground. Let us review the steps which lie ahead of us.

There is, in the first place, an added task of analysis which is so far undone. This is the explanation of the relations between money expenditure and money income, and of both of these to the general level of money prices—an analysis which involves more broadly the explanation of the level of employment. In our preceding discussion we have "explained" nothing along this line. Instead we have simply postulated various possible relationships of income to expenditure—resulting in a rising, falling, or stable flow of payments—and have traced what might happen to employment and output in the several possible situa-

[10] See Hicks, *op. cit.*, Chap. 8 and Appendix to Chap. 8, for a formal treatment.

tions. We have not *explained* under what conditions the flow of money payments will rise, fall, or remain stable, or the manner in which this flow of payments, the level of prices, and the volume of employment may be interrelated and may tend to seek some mutual equilibrium. A theory of income and employment should explain this if it is to provide an adequate interpretation of a free-enterprise economy. The development of such a theory, however, is a sufficiently large task to require a separate volume. We will be limited in this work to a brief treatment of the theory of income and employment, particularly in connection with the discussion of capital and interest.

Several further tasks which cannot legitimately be left to further volumes involve elaborations of the factor-pricing analysis which was begun in this chapter. First, special attention must be given to each of the principal factors of production, as regards the considerations influencing its supply price and its particular role in the processes of production. For the sake of simplicity we have so far dealt with the idea of "factors *A*, *B*, and *C*," as if productive factors were any three basic inanimate commodities, which firms bought and combined into more finished commodities. We thus abstracted from any potentially significant differences between such commodities and actual factors of production. In fact, each factor has special characteristics which makes the determination of its price and of the quantity of it bought by firms somewhat different from that for commodities. These distinguishing characteristics, and their importance for economic behavior, must be investigated. We must examine the special character and the conditions of supply for each factor, the role of each factor in production, and the manner in which this role may influence the character of the demand for the factor. It will appear, for example, that the parallel between either labor or capital on the one hand and a "basic commodity" on the other is perilous indeed, and that the special character of such factors of production clearly conditions the nature of the economic process.

A second extension of our analysis must involve its modification to take into account the fact that factor markets, as well as other markets influencing the demand for factors, are not purely competitive. We have so far traced factor-price deter-

mination on the assumptions (1) that all firms buying factors sold their outputs in purely competitive markets, (2) that the number of buyers in each factor market was very large, (3) that the number of sellers in each factor market was very large, and (4) that for each factor there was a single, homogeneous, economy-wide market. As a first step, these assumptions must be altered to accord with the facts of the modern economy. That is, (1) most firms sell their outputs in oligopolistic markets, with some corresponding effect on their demand for factors; (2) the number of buyers in a market for factors is often small, introducing a degree of monopsony into the factor-pricing process; (3) the number of sellers in some factor markets, especially for unionized labor and for capital, may be few, introducing an element of monopoly into the process; and (4) instead of a single market for each factor, there is an overlapping series of submarkets for various parts of the factors—for segments which are imperfect substitutes for each other because of differences in type, quality, or geographical location. After altering our assumptions to accord with fact, we must see in what wise the processes of factor pricing and income distribution are altered by degrees of monopsony and monopoly and by differentiation among subcategories of factors. This may conveniently be done as we give detailed attention to each factor in turn.

A final task involves the further examination of profits. In the preceding analysis of income distribution in purely competitive equilibrium, profits appear as a distributive share which has a tendency to vanish—or which emerges only as a result of temporary or chronic disequilibrium in the economy. Although these observations on profit may be correct for a competitive economy approaching general equilibrium, profits must also be examined in the context of a world with many elements of monopoly and with a continued tendency to dynamic change. We must therefore return to the problem of profits within a broader and more realistic framework of analysis.

In the light of the remaining task ahead of us, succeeding chapters will concern themselves with: (1) capital, interest, and money—the markets for them, and their relation to the productive process; (2) the markets for labor, and the determination of wages and of the employment of labor; (3) the markets

for land and resources, and the determination of rents; and (4) profits as a distributive share in a dynamic and imperfectly competitive economy. We must also aim at some summary conclusions on the general equilibrium of the economy, and on the determination of the level of income and employment.

SUPPLEMENTARY READINGS

JOSEPH A. SCHUMPETER, *The Theory of Economic Development*, Cambridge, Mass.: Harvard University Press, 1934.

J. R. HICKS, *Value and Capital*, Part II.

OSCAR LANGE AND F. M. TAYLOR, *On the Economic Theory of Socialism*, Minneapolis: University of Minnesota Press, 1938.

JOHN MAYNARD KEYNES, *The General Theory of Employment, Interest, and Money*, New York: Harcourt, Brace and Company, 1936, Chap. 3.

J. R. HICKS AND A. G. HART, *The Social Framework of the American Economy*, New York: Oxford University Press, 1945.

CHAPTER 11

CAPITAL GOODS AND INVESTMENT

The simplified treatment of income distribution so far has established, subject to drastic simplifying assumptions, certain principles governing the volume of employment and the absolute and relative prices of factors of production. It would appear that the level of employment depends ultimately upon the mutual adjustment between the level of flow of money income and the level of money factor prices. The relative prices of various factors of production, when supplied in given amounts and under competitive conditions, depend principally upon the technical substitution relations among them; the price of each factor tends to assume a ratio to any other factor price which is the inverse of the marginal rate of substitution between them throughout the economy. These principles are generally correct, and in a general way applicable to a real economy employing labor, land, and capital. But certain simplifying assumptions which we have made are quite arbitrary, and our conclusions require re-examination in the light of the special characteristics of various factors of production.

We have so far taken the view that each of several factors of production is in essential respects similar to some basic inanimate raw material, which is given in quantity to the economy without being produced, and upon which all firms draw for raw material for production. The supply conditions for such "commodities" have not been investigated; rather we have simply assumed

that each factor might have various alternative "supply curves" relating the amount supplied to either the money or the real price which buyers offer. Special considerations influencing the supply of labor, land, and capital have thus been overlooked, as we referred to three imaginary or arbitrary factors, *A*, *B*, and *C*. We have also assumed that each of these several factors is strictly an independent substitute for the others—that any firm, or the economy as a whole, may simply use less of one and more of another, and that none of any factor "*A*" is used to produce other factors "*B*" or "*C*." Neither of these assumptions corresponds closely enough to reality that it can be left unmodified as the basis for analysis.

THE CHARACTER OF CAPITAL

Capital, as a factor of production, requires particular attention from this standpoint, both because it is not strictly independent of other factors, and because it has rather special and peculiar conditions of supply. To understand its peculiar character, we must distinguish among three conceptions: (1) capital goods as that stock of "instruments of production" already on hand at any current moment which serves as the temporal reference point for analysis; (2) capital goods as commodities to be produced following this point in time for either replacement of or addition to the existing capital stock; and (3) capital as "investable funds," which have been invested in existing capital goods and which may be recovered and used for reinvestment (replacement) as these goods wear out, which may be invested in additions to the stock of capital goods, as these are produced, and which also may be and are used to purchase any earning assets already in existence, including old securities, land, etc.

At any current time in a developed industrial community, we have inherited a very large stock of capital goods, which on the average are sufficiently durable to last for some additional time interval. These include factory buildings, machinery, tools, stock piles of raw materials, and so forth. Looking at our analytical problem from this current date, such capital goods are temporarily fixed factors, and currently just as given in amount as the supply of agricultural land or of mineral resources. As such,

they may be considered, for a short run lasting until they wear out or are used up, as simple basic commodities which provide services for use in production, and as independent substitutes for land and labor. Their services are subject to pricing on this basis, for whatever they may be worth in terms of their substitution relations with other factors. For such capital goods—and necessarily for the short run, since they are exhaustible—the view that we have an additional independent factor is not greatly amiss. The historical fact that these existing capital goods represent in an ultimate sense the embodiment of the services of labor and land—and this alone—is interesting and perhaps instructive, but it is not strictly relevant to the current problem of analysis. The capital goods are here, and that is that. Neither, as a matter of fact, is their historical cost, in terms either of money or of land and labor services embodied, a relevant magnitude, since this cost is "sunk" and should not influence the current use of these existing capital goods. Existing capital goods are a sort of "quasi land," and their earnings correspondingly "quasi rents."

A certain amount of funds have in the past presumably been invested in these goods, and currently "remain invested." Correspondingly, a part of the services of this stock of capital goods are yet to be rendered, and will earn incomes which will be available to reward the investors with "interest," return of "principal," or both. The amount of this past investment of funds is not strictly relevant to current analysis, but it is important that the investors as a class will retain certain claims on the earnings of the capital goods.[1] They will have liens for payment of interest and return of principal if they are creditors, or equity and dividend rights if they are shareholders. These contractual claims, arising from past investment, will influence the current distribution of income so long as they remain in effect—that is, until they are revised or until the funds in question are disinvested. For the short run, therefore, we have a given stock of capital goods, which temporarily assume the character of inde-

[1] It is also relevant that these claims on the future earnings of existing capital goods, in the form either of old securities or the direct title to the old assets, are currently salable and constitute one portion of the current aggregate demand for liquid funds, if the interest rate is such as to induce the asset holders to sell. We will return to this matter in Chapter 12.

pendent productive factors, and a system of claims to their earnings, resulting from the past provision of funds, which currently influences the distribution of income.

This simplified view, however, is accurate only for a short period following any given beginning point of analysis. As time passes (or as its passage is anticipated for purposes of analysis) various existing capital goods, and ultimately all of them, will wear out or otherwise be exhausted, thus becoming subject to replacement with newly produced capital goods. At the same time, there is the possibility of expanding the capital stock with additional items. As we contemplate the passage of time, therefore, it becomes clear that capital goods are in the long run not independent factors but commodities to be currently produced (or not) as the outputs of labor, of land, and of the initial stock of capital goods. Taking any given beginning point, with some existing capital stock, the economy has the choice (1) of producing capital goods at a slower rate than those in existence wear out, thus reducing the capital stock, disinvesting in the net, and freeing some resources for the production of other than capital goods, (2) of producing capital goods just at the rate that those in existence wear out, thus maintaining the capital stock, undertaking no net investment or disinvestment, and keeping a constant proportion of resources employed in producing (actually reproducing) capital goods, or (3) producing capital goods at a more rapid rate than required for replacement, thus adding to the stock of capital goods, undertaking *net investment*, and devoting some proportion of productive resources continually to this process of addition.

Whatever the choice turns out to be, the use of capital goods over time does not involve the introduction of an independent substitute for labor and land. It involves, rather, the use of the services of labor and land (and of any initially existing capital stock) in a particular way—to produce goods for further use in production. In effect, the use of capital goods, rather than substituting for the services of labor and land, involves a special routing of the flow of these productive services through the process of production—a demand for these productive services in a special and generally indirect form. By using more or less

capital goods in combination with labor and land, the economy is essentially routing a smaller or larger proportion of the basic productive services of labor and land through the production of capital goods and thence to the production of final goods for consumption.

In the shoe industry, for example, shoe machinery is used in production, and as it wears out it is replaced. As it is maintained, it replaces or is substituted for large amounts of manual labor which would otherwise be required in the shoe factories. But at the same time the use and regular replacement of such machinery involves the continued indirect employment of labor, land, and other capital goods to produce shoe machines. For the economy, the use of capital goods (shoe machines) does not involve a substitution for the services of other factors but a more "roundabout" use of their services. Depending upon whether the capital stock is being used down, just maintained, or augmented, the "degree of roundaboutness" is decreased, maintained, or increased.

This last idea has sometimes been expressed by introducing the idea of the "structure of production," which refers generally to the pattern in which resources are employed in the economy as a whole—how many directly in the production of consumer goods, how many in the production of capital goods used in producing consumer goods, how many in the production of capital goods used in producing other capital goods, etc. The structure of production then is "elongated," generally in the direction of increasing the proportion of resources employed in the latter two categories, when more capital goods are used, or is "shortened" if fewer capital goods are used in production. Although "lengthening" and "shortening" are oversimplified one-dimensional concepts as applied to an essentially multidimensional structure of production, the general validity of the idea is clear. It is quite clear that in industrial societies since 1800, for example, there has been a progressive and rapid "lengthening" of the structure of production, as an increasing proportion of resources has been devoted to the production of capital goods.

For purposes of other than arbitrary short-run analysis, in sum, capital goods must be viewed as a part of the output of

available resources, and not as independent substitutes for these resources. The use of capital goods involves the indirect use of productive services otherwise available for more direct use. The analytical problem with respect to the use of capital goods is thus not that of simple substitution of one factor for others, but one of the determination of the relation of the demand for productive services devoted to making capital goods to the demand for these services devoted to making other goods—of the proportion of productive services devoted to indirect as opposed to direct use.

This would be a complete problem in itself even if no additional complications were involved, although it would not introduce an additional "factor of production" or an additional distributive share of income. An additional complication is introduced, however, by the fact that investable funds are required to finance the acquisition and retention of capital goods. Capital goods are, on the average, "durable," or are acquired for use over some finite interval following purchase.[2] They must be paid for when acquired, but they are used and provide an earning from use with some average delay after acquisition. It follows that firms in the economy require funds to invest in capital goods. When new capital goods are added, additional funds must be secured to finance their acquisition. As existing capital goods are maintained by replacement, the initial funds (or their equivalent) must remain invested or be reinvested to maintain the capital goods in use. The demand for capital goods thus carries with it a corresponding demand for investable funds. There is an additional service involved in the use of capital goods—the service of providing funds to finance the acquisition and holding of these goods.[3]

[2] Strictly, there is a time lag between the date of acquisition of the goods and the sale of the output to which the use of their services gives rise.

[3] All goods require funds. Investable funds, however, are funds other than those spent to add to immediate consumption and thus require that the persons who supply the funds to buy the goods receive some reward other than the consumption of the goods purchased. As noted, investable funds are not limited in their use to newly produced capital goods but may be used generally to purchase title to future income streams, as from land. The demand for investable funds is thus not limited in origin to the demand for new and replacement capital goods currently produced.

Such funds may be provided either on loan, by creditors, or as investment, by owners or in return for equity. In either case, there may be a charge—*i.e.*, *interest*—for the provision of such funds, and thus a corresponding distributive share of income paid to the creditors or owners. This in general is a charge in addition to the original purchase price of the capital goods financed—and thus in addition to the prices paid to the productive services which are embodied in the capital goods. It follows that the capital goods when put to use must be expected to "earn" not only their original purchase price but also a premium sufficient to pay whatever interest charge there is on the invested money. It is also true that persons or institutions (such as banks) which are able to supply funds may be able to earn a share of income in return for making funds available.

At any current time, as we have pointed out, many past investments of funds will have been made, and will be reflected in various contractual claims which creditors or investors hold on the income of various firms. As we proceed forward from this point, the maintenance of existing capital goods involves a continually recurring demand for sufficient reinvestment of these funds to finance replacement. Additions to capital goods would involve demands for additional new funds for net investment. In either event, a continued demand for invested money is implied in the use of capital goods. Only in the event of net additions to capital goods, however, will a net new supply of investable funds be required.

With investable funds introduced as an essential part of the capital-using process, we are now able to frame the analytical problem surrounding the use of capital goods. First, what determines the demand for capital goods (for the use of resources to produce capital goods) and thus also the demand for investable funds? Second, what determines the supply price of capital goods, as purchased from those who produce them? Third, what determines the supply price of—*i.e.*, the interest rate on—these investable funds which are used to finance the purchase of capital goods? And finally, how do this demand condition and these two conditions of supply interact to determine the amount of capital goods used, the rates of gross and net investment through

time, and the size of the income stream going to investors in capital goods as interest?

THE DEMAND FOR CAPITAL GOODS

The origin of the demands for capital goods and the determination of their supply prices are really parts of one question and may be considered together. The early economic theorists who probed the problems of capital perceived the essential facts when they observed that capital goods are used because they increase productivity—because labor and land when used in roundabout fashion to produce capital goods are ultimately more productive of final consumer goods than when used directly. They pointed to the obvious superiority of introducing roundabout, or capital-goods-using, methods of production. As a general principle this is quite correct, but it requires further examination. We may attempt to penetrate the issue further by inquiring (1) why capital goods will be demanded, and in what amount, in a given or stationary "economic situation," with given supplies and employment of labor and land, a given set of final or consumer products, given consumer tastes, and a given state of technical knowledge, and (2) how the demand for capital will be affected by economic change affecting amounts of labor and land, available products, consumer tastes, and knowledge of productive techniques.

Let us first take an economy with a given supply of labor and land at full employment, a given list of consumer products to be produced, a given state of consumer choice as among these goods, and a given state of knowledge concerning available production techniques. Supposing that no changes occur in any of these basic conditions, what will determine the demand for capital goods and for investable funds? This demand originates in a group of firms producing an array of consumer goods to satisfy a given family of demands for these goods. If these firms elect to use any capital goods at all in production, it must be because they can thus improve the ratio of output to cost—that they can produce the same output with lower cost, or a larger output without a proportionately larger cost. There must

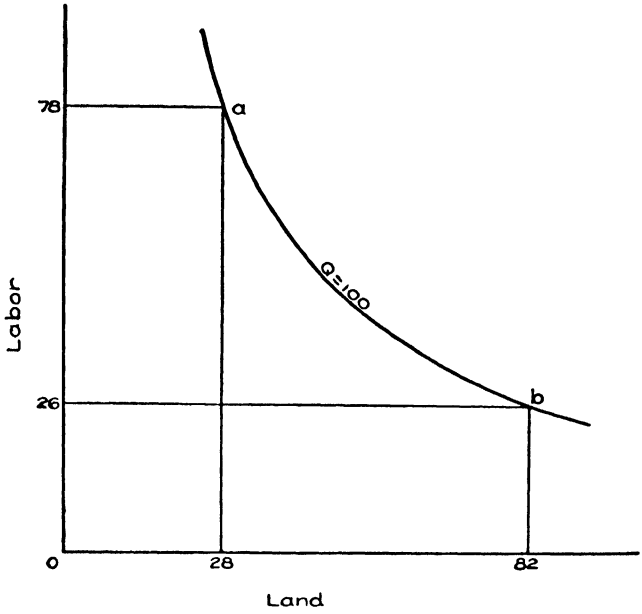

Figure 48

in effect be some virtual advantage to the use of capital goods in production which is not fully offset by the price for the services of capital goods.

The character of the virtual advantage of capital goods presumably is derived from the general law of substitution between productive services. As between labor and land, for example, we have seen that a given output may be produced with various proportions of the two factors, but that as any one factor is increased at the expense of the other, its marginal rate of substitution for the other factor increases. Thus the isoquant in Figure 48 shows the various proportions of labor and land which may be used to produce a given quantity of product. The quantity 100 of product may be produced with all the indicated proportions of labor and land, including, at point *a*, 78 labor and 28 land, and, at point *b*, 26 labor and 82 land. But as the amount of labor is increased (moving, for example, from point *b* to point *a*, the marginal rate of substitution of labor for land in-

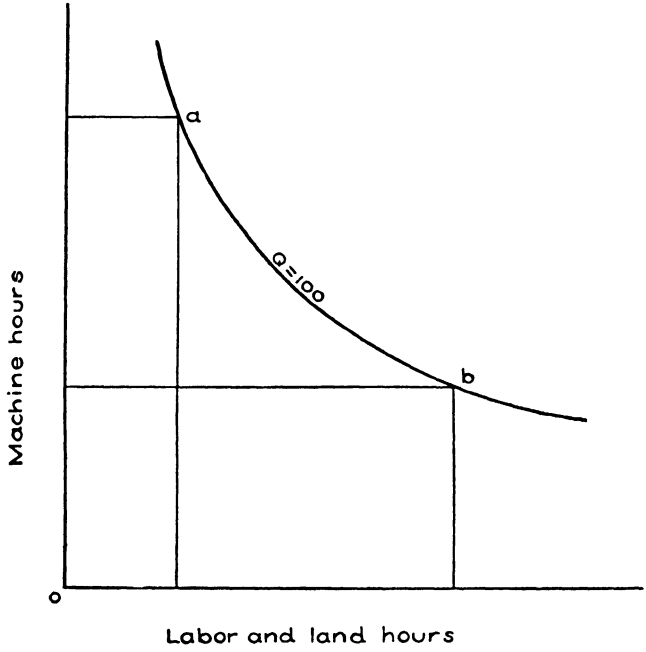

Figure 49

creases progressively. At point *b*, a very little labor will replace a unit of land; as we approach *a*, more and more labor is required. We have also seen that the firm in competition will use these factors in such proportion that the marginal rate of substitution of labor for land is equal to the ratio of price of land to that of labor.

Precisely the same principle applies to substitution between capital goods services and other factors. The substitution relationship between machine hours of service and "labor-and-land" hours of service[4] for a firm should follow a similar pattern, as shown in Figure 49. Machine hours may be substituted for labor-and-land hours at a virtually very advantageous ratio at point *b* or to the right of point *b*. That is, if the firm is at the outset using little or no machinery, it can replace many labor-

[4] Labor and land may be provisionally viewed as a single "composite" factor against which capital goods services may be substituted.

and-land hours with a few machine hours. But as it successively adds machine hours (moving toward point *a*) the marginal rate of substitution of machine for labor hours progressively increases.

Does this relationship make it inevitable that the firm will employ at least some machinery—*i.e.*, substitute capital goods for labor and land—and if so how much will it employ? The answer, of course, depends on the relative prices of labor-and-land hours and of machine hours. If the price of machine hours is relatively high (say three times that of labor-and-land hours) the firm would employ relatively few machine hours. If the price of machine hours were, instead, relatively low, they would be substituted for labor-and-land hours up to the point where the marginal rate of substitution was in balance with new price ratio.

Why are the services of capital goods—*e.g.*, machine hours—employed at all? Very evidently because the price of such services is sufficiently low (relative to the price of other factors) that they may be substituted to advantage for the services of other factors. The reason for this advantageous price ratio, in turn, is found in part in the determination of the price of capital goods. The capital goods which supply machine hours or similar services are commodities produced through the use of labor, of land, and of the services of other capital goods. Their cost of production, therefore, is made up of wage rates for labor and of rents for land, plus, for embodied capital-goods services, a charge which tends in the long run to equal the replacement cost of these services, ultimately in terms of wages and rents. In effect, the costs of production of capital goods tend to approximate the prices of the labor and land used either directly or indirectly in their production.[5] Under competitive conditions the prices of capital goods tend to equal these production costs; if the capital goods are sold in a monopolistic market, their prices may tend

[5] This may be accepted as a simplified first approximation for the purposes of argument. Where time is consumed in making a given capital good, it must earn interest from the dates of the first inputs of factor services, not simply from the date of its completion. Then its cost at the date of delivery already includes an accumulated interest charge on earlier inputs. Recognition of this complexity would not modify the general tenor of the following argument.

to exceed their production costs somewhat. But in any event, the prices of capital goods, and thus of machine hours, tend to be governed at least roughly by the labor and land costs ultimately embodied in them.

The *ratio* of the price of capital-goods services to that of labor or land thus tends toward a ratio between the price of labor and land indirectly used (and representing the cost of capital goods) and the price of labor and land directly used, the wage and rent rates being roughly the same in both cases. With this general limit on this price ratio, some capital goods will be used so long as a "labor hour" used (for example, in making machinery to make shoes) contributes more to ultimate output than a labor hour used directly to make shoes. Such superiority of "roundabout" labor and land is generally found up to a certain point as capital goods are substituted for other productive services. The profitability of substituting capital goods for direct labor and land thus rests on the two facts (1) that given amounts of labor and land are more productive up to a point if used to make capital goods for production than if used directly in production, and (2) that the price of capital goods thus produced ultimately tends to correspond to the cost of labor and land employed, directly or indirectly, in making them. Under this condition, it is evident that capital goods might be substituted for other factors in considerable degree before the rate of substitution of capital-goods services for other factor services came into balance with the governing price ratio.[6]

Of course, the individual firm which considers acquiring a capital good does not necessarily inquire why a machine costs what it does. But taking the price of machine hours as given, together with those of labor and land, it will consider the substitution relation (as illustrated in Figure 49) between machine hours and other factor services, and undertake substitution of capital for other factors up to the point where the marginal rate of substitution of capital for labor is in balance with the price

[6] See F. A. von Hayek, "The Mythology of Capital," *Readings in the Theory of Income Distribution*, edited by W. Fellner and B. F. Haley (Philadelphia: The Blakiston Company, 1946), Chap. 20, for a statement of doctrine concerning the yield of capital and related matters.

ratios. The substitution relations between capital and labor and land, for example, might appear as follows in a typical instance:

100 Units of Product Producible with the Following Combinations

Machine hours	Labor-and-land hours
10	5
9	5.75
8	6.62
7	7.62
6	8.87
5	10.37
4	12

For successive discrete movements from 4 to 5, 5 to 6, etc., of machine hours, the marginal rate of substitution is successively 1:1.63; 1:1.5; 1:1.25; 1:1; 1:0.87; 1:0.75. If the price of a machine hour were $1 and that of a labor-and-land hour also $1, so that the price ratio were 1:1, the firm would substitute capital up to the point where the marginal rate of substitution became 1:1—to where between 7 and 8 machine hours were employed per 100 units of output, since at this point cost would be minimized. If, on the other hand, machine hours were $1.25 whereas labor-and-land hours were $1, the price ratio being 1:25 to 1, the firm would substitute only up to the combination where the rate of substitution of machine hours for other services was 1:1.25 or in the neighborhood of 6 to 7 machine hours, since this would now minimize cost. Given the cost of capital goods, each firm's employment of these goods relative to other factors is strictly determinate according to the usual principles.

The preceding discussion outlines the basis of the firm's, and thus the economy's, tendency to demand capital goods, and therefore to give rise to such an organization of the structure of production as will supply the capital goods they demand. Two remarks should be added. First, under given fixed conditions of technique, tastes, products, and supplies of labor and land, and with a given ratio of capital-goods prices to their costs, firms would tend to seek a determinate balance in the use of capital-goods services—a balance which is defined in terms of a ratio or *proportion* of capital-goods services to other services. The *absolute amount* of capital-goods services employed consistent with

this ratio will depend also on the level of employment of labor and land for the economy as a whole—the closer that the economy approaches full employment, the more capital-goods services will be required. So far we have not demonstrated what this level of employment will be, but only the tendency of firms toward a definite demand for capital-goods services at each possible level of output and employment. At any *given* level of employment, however, whether full or otherwise, and in the generally stationary situation postulated, the absolute amount of capital-goods services required is strictly determinate. The economy will reach an equilibrium in which it will employ a given constant quantity of "machine hours" in each successive period of time.

The second point is that if this is true in the indicated situation, the economy's demand for "machines"—for the capital goods supplying productive services—will be strictly satiable. Corresponding to the requirement for a certain number of "machine hours" will be a certain number of "machines," or stock of capital goods. Once the firms of the economy have acquired this stock of capital goods, they will not wish to enlarge it but only to maintain it. Thereafter they will not purchase *additional* machines but only as many as are required to replace items in the equilibrium stock as they wear out. The demand for *additional* capital goods is fully satiated—only *replacement demand* remains once equilibrium in a stationary situation is struck.

Correspondingly, additional funds for net investment are not required once an equilibrium stock of capital goods is acquired, since, to finance replacements, it will only be necessary to retain previously invested funds for reinvestment. A continuing demand for new investable funds is thus unlikely to emerge from a stationary economic situation, for it is only successive additions to the stock of capital goods which will continually refresh such a demand.

The relative stability, in a stationary situation, of the demand for capital-goods services, together with the tendency of the demand for additional capital goods to approach zero, is in part accounted for by the fact that the price of capital goods (and of their services) is not a free price which moves independently of

the prices of other factors. As we have seen, the price of capital goods is based on wages and rents paid to produce them, as well as on the wage and rent replacement cost of other capital-goods services used to produce them. As a result, capital-goods prices will not tend to vary if wages and rents are generally stable (except so far as price-cost relations change) and they will tend to respond directly to changes in wages and rents. The *ratio* of capital-goods prices to labor and land prices thus tends to remain relatively stable (with given techniques), and the stable long-run balance referred to is unlikely to be seriously upset by factor-price variations.

INTEREST COST AND THE USE OF CAPITAL GOODS

We have so far examined the source of the demand for capital goods in abstraction from the interest cost of investable funds, showing to what extent firms would employ capital goods if in effect the interest charge were zero. It must now be recognized that when firms acquire capital goods and thus require investable funds they must pay or impute an interest charge on these funds. This interest charge constitutes an addition to the cost per machine hour. It thus tends to restrict somewhat the degree to which capital-goods services are substituted for others.

In the preceding example, we supposed the cost per machine hour, exclusive of interest, to be \$1. This cost is presumably calculated as the sum of the original cost of the machine (which must be recovered as depreciation) plus the total cost of operation and maintenance, divided by the number of machine hours rendered over the economic life of the machine. If, in this example, the unit machine (rendering 1 machine hour per hour) originally cost \$5000, would provide 3000 hours per year, would last 5 years, and would cost \$2000 per year to operate and maintain, the cost per unit of service, U, could be calculated as follows:

$$U = \frac{\$5000 + 5(\$2000)}{5(3000)} = \$1.$$

Generally, letting C stand for original cost, O for operating cost, and S for service units of service, and o for the date of

acquisition, n for the end of the economic life of the machine, and t for time, we have:

$$U = \frac{C_o + \sum_{t=o}^{t=n} O}{\sum_{t=o}^{t=n} S}.$$

(The scrap value of the machine is assumed to be zero.)

This is the derivation of our cost per machine hour, neglecting interest cost. Now if the firm must pay interest on investment, the interest cost must be added to the cost per machine hour. With a $5000 machine, average annual interest would equal (*in rough approximation only*) the interest rate times one half of the necessary investment, since over the life of the machine on the average about half of the original cost will be invested, the remainder being progressively recovered as the machine is used. Then, if the interest charge on money at the market rate were 4 percent, the *annual* interest charge would equal $(5000/2) \times .04$ or $100, and the machine-hour cost inclusive of interest would be calculated as:

$$U = \frac{\$5000 + 5(\$2000) + 5(\$100)}{5(3000)} = \$1.03\tfrac{1}{3}.$$

The total cost per machine hour thus becomes $1.03⅓, and in substituting capital goods for labor and land the firm will not go quite so far. If labor and land cost $1 per hour, for example, it would proceed only to the point where the marginal rate of substitution of machine for land and land hours was 1:1.03⅓.

Unlike the basic price of capital goods, the rate of interest is potentially a free and independent price and may move independently of wages and rents. It is then evident that the demand by firms for the services of capital goods will be somewhat influenced by the rate of interest—a higher rate of interest will somewhat restrict the degree to which capital is employed, and a lower rate encourage it. Thus a rate of interest of 8 percent in the preceding example would raise the price per machine hour to $1.06⅔, and a rate of 2 percent would drop it to $1.01⅔. A rate of interest of zero would, of course, drop the price to $1. As the interest rate is progressively lower, larger amounts of

capital would tend to be employed, and correspondingly larger amounts of investable funds demanded. Thus, in the example on page 323, we have seen that with a zero interest rate the price of machine services might be $1 per hour, labor-and-land services also being $1 per hour. At this rate, between 7 and 8 machine hours would be employed per hour to produce 100 units of output, and thus 7 to 8 machines (let us say 7½). Investment, at $5000 per machine, would be $37,500.[7] Corresponding to rates of interest of 2, 4, and 8 percent, however, we would use fewer machine hours, let us say respectively 7.0, 6.8, and 6.7. For the firm in question, therefore, initial investment in machines per 100 units of output would vary as follows (at $5000 per machine) in response to changes in the interest rate:

Interest rate	Investment
.08	$33,500
.04	34,000
.02	35,000
.00	37,500

From these extremely rough calculations we derive one significant impression—the firm's (and the economy's) demand for capital goods and for investable funds varies in response to the interest rate, after the fashion of any conventional demand schedule. With a given interest rate, as postulated in the preceding section, the economy in a stationary situation will tend to employ one determinate amount of capital goods, and a corresponding definite amount of investable funds for this purpose. For various alternative interest rates, it will have a definite schedule of demands for capital goods and for investable funds, negatively related to the rate of interest. This schedule is negatively sloped because as the interest rate falls, the price (including interest) of capital-goods services tends to be reduced, capital-goods services are further substituted for other factor services, the output corresponding to any given level of employment is increased, and the demands for these goods and for investable funds are correspondingly increased. This is true even though employment for the economy is not influenced by the

[7] That is, $37,500 new, or $18,750 in an average state of 50 percent depreciation.

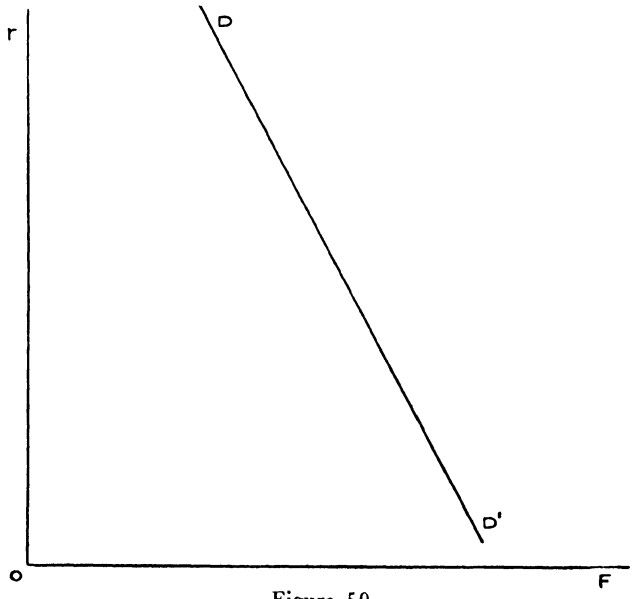

Figure 50

reduction of the interest rate. If an increase in employment does result, the effect of interest-rate reductions on the demand for investable funds will, of course, be greater. (It will be noted that a zero rate of interest here refers to a situation where principal is loaned subject to no interest charge, but must be repaid at the end of a finite time interval, ordinarily corresponding to the useful life of the asset in which the funds are invested. The loan of funds at zero interest with indefinite renewal privileges would amount to giving money away, and is not considered here.)

The effect of changes in the interest rate on the level of employment and output will be considered later in this chapter. For the moment, it is sufficient to observe that at any time there is some negatively inclined economy-wide demand curve for investable funds, showing the response of amounts of funds required (F) to changes in the interest rate (r), as in Figure 50. A demand schedule which shows the amounts of investable funds demanded in the economy at each of a number of alternative rates assumes a given position corresponding to:

1. Given techniques of production.
2. A given list of available consumer goods for production.
3. A given state of consumer choices as among these products.
4. A given supply of labor and of land.
5. A given initially going level of employment.
6. A given ratio of capital goods prices to other factor prices and commodity prices.

Taking all of these things as given (and constant for purposes of argument) we find a determinate "real" demand schedule for capital goods, indicating the amounts of capital goods demanded by firms at each rate of interest. This real demand is translated into an actual demand for dollars of funds when in addition we know also the general money price level, which will determine the total money payment corresponding to each real demand. We may thus conclude that, given the first six determinants, there is at any time a determinate real demand schedule for investment in capital goods, and that given these *and* the money price level, there is a corresponding demand schedule for dollars of investable funds.

This is the demand for the "money factor" which contributes to capitalistic production. In any stationary situation of techniques, products, employment, price level, etc., this curve has a given position. It shows that at the rate of interest r, Q dollars of investable funds will be required to finance production; that at the rate of interest r_1, Q_1 dollars of investable funds will be required; etc., as in Figure 51.

It must be re-emphasized, however, that this demand for funds is strictly satiable at any specific interest rate.[8] With a given interest rate r, for example, Q dollars of investable funds will be required to finance the purchase of the corresponding equilibrium stock of capital goods. As these goods are acquired, this amount of funds will have to be secured from investors. But the maintenance, by periodic replacement, of this stock of capital goods will require no *additional* funds. Once the capital goods are on hand, the receipts from the sales of the outputs they produce will in equilibrium provide sufficient funds to pay for replacement—that is, the initial investment of funds will be

[8] See J. M. Keynes, *op. cit.*, Chap. 16.

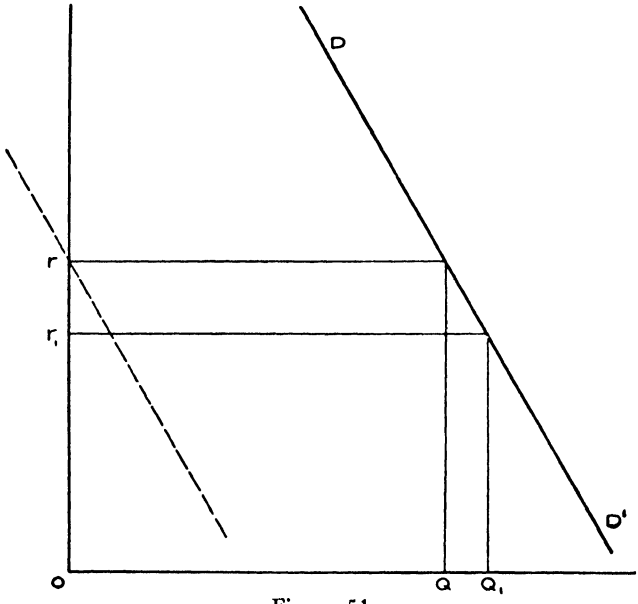

Figure 51

regularly recovered from sales (as depreciation costs) as the capital goods are used up. It follows that to maintain a given equilibrium level of investment in capital goods, once it has been reached, requires no further *net investment* of funds, but only regular *reinvestment* of previously invested funds as they are recovered from production. In a stationary economy there is thus not a perpetual demand for *additional* investable funds for capital goods. Once firms have acquired the optimum amount of capital goods, they will simply replace them, and there will be no net new investment opportunities. Capital goods will continue to be produced at a constant average rate to replace the existing stock as it wears out, and will continue to be used, but no net new funds will be required to finance replacement. Simple abstinence from withdrawing past investments of funds will suffice.

The demand schedule we have attributed to a given stationary situation thus shows the total amounts of investable funds required "for all time," not the additional amounts re-

quired in successive time periods. It is a gross accumulated demand for investable funds, and it is only as such that it maintains its indicated position as the economy passes from one time period to another. As soon as this demand has been once satisfied at any rate of interest r—as soon, that is, as Q dollars have been invested—the net investment demand at that rate drops to zero (as shown by the dotted line in Figure 51) and remains there. When equilibrium is reached at a specific rate of interest, the entire indicated demand for funds,[9] DD', is thus reflected in reinvestment demand, and in equilibrium can be financed entirely from the sales receipts of output. In a stationary equilibrium situation, therefore, the demand for new investable funds, as supplied for example by savings, declines to zero, and can be stimulated only so far as the rate of interest can be reduced.

Why is it significant that in a stationary situation the demand for funds is satiable, in the sense that it develops entirely into a demand for reinvestment of funds recovered as a part of the sales receipts of firms using capital goods? Because, in such a stationary equilibrium, the full flow of all individual net incomes—wages, rents, interest, and profits (if any)—finds no outlet for expenditure except on consumption goods. When an economy is acquiring additional capital goods, and thus demanding new investable funds, a part of current net incomes may be diverted as savings to supply these funds. A share of net incomes is thus spent on capital goods, and does not have to add to current consumption to be spent at all. But it is spent in one way or another, recirculates, and becomes income again in the succeeding time period. When this succeeding period has found an equilibrium stock of capital goods, however, none of the current total of net incomes will be demanded for investment or spent on capital goods, since the purchase of these for replacement will be financed by enterprises out of the amounts individuals spend on the consumption goods. Consequently, all net incomes must turn to consumption expenditure to be spent at all, or, in effect, there

[9] Excepting any net demand for funds which may result from the demand for consumer loans, the offer of land for sale, or the offer of other earning assets already in existence. In stationary equilibrium, this net demand would also tend to drop to zero.

will no longer be any place for individual net saving.[10] The appropriate relation of consumption spending to income thus changes as we move from a growing to a stationary economy, and from a process of net investment to one of simple reinvestment. And it may be possible that individuals will be slow or reluctant to modify their spending-saving habits accordingly.

INVESTABLE FUNDS AND GENERAL EQUILIBRIUM

It may be useful at this point to connect the argument of this chapter with that of the preceding one. In Chapter 10 we demonstrated how a competitive general equilibrium might be struck in an economy which drew upon three inanimate and independent factors, *A*, *B*, and *C*, where each factor was available without production cost in a certain quantity or subject to a certain given supply schedule. The attainment of a balance of factor employment, factor prices, and commodity outputs and prices was demonstrated, subject to the attainment of stability in the flow of money income through the economy. Now suppose that instead of three factors we have two primary factors, *A* and *B* (corresponding to land and labor) each available to the economy along a given supply schedule. In addition we have a third, namely investable funds, which, let us say, is available in perfectly elastic supply at a given interest rate. (We have yet to examine the determination of the supply price of investable funds.) Finally, we have the possibility of using investable funds to convert factors *A* and *B* into durable capital goods, with a corresponding advantage, up to a margin, in over-all efficiency or cost. At any current time, long after such a process of conversion was begun, we have on hand a large stock of such capital goods, already integrated into the structure of production, and requiring periodic replacement. We are viewing a stationary economy—stationary in final products, techniques, consumer tastes, and supplies and employment of both factors *A*

[10] Unless the offer of old assets (including land) and the demand for consumer loans provides some net demand for investable funds. In stationary equilibrium such net demand would ordinarily tend to be zero—hence we neglect it here.

and *B*. How now is our analysis modified by recognizing that factor *C* is investable funds?

Previously it was pointed out that with three independent factors, *A*, *B*, and *C*, an equilibrium level of employment and output could be struck, with the level of money factor prices, that of money commodity prices, and that of money purchasing power in balance. (Full employment or less than full employment are conceivable under alternative assumptions.) In this balance, the ratios of the prices of *A*, *B*, and *C* would be in balance with the marginal rates of substitution among them. If now capital—*i.e.*, investable funds—is introduced as the third factor, our determinants are altered in this wise. Factors *A* and *B*, arbitrarily representing labor and land, have given supply schedules of presumably positive slope. Investable funds have a perfectly elastic supply curve at some assumed interest rate. The economy of firms, beginning with some initial flow of real purchasing power (which we will temporarily assume to be self-maintaining) draws upon these three factors. As it does so, capital goods will tend to be acquired up to some equilibrium level—or held at that level if it has already been reached—a level such that the marginal rate of substitution of capital-goods services for other factor services is equal to the ratio of the price of these other services to that of capital-goods services, inclusive of the interest charge. This will result in an organization of production such that a given proportion of the services of labor and land are continually devoted to or demanded in the production of capital goods, for replacement of the equilibrium stock thereof, and that a corresponding amount of investable funds is held in investment to finance the retention of the capital stock.

Net new investment of funds will be required, however, only while the economy is moving toward the equilibrium stock of capital goods. Once this stock is reached, the rate of net investment necessarily drops to zero. Some downward adjustment of the interest rate may therefore take place, but wherever it comes to rest, net investment again drops to zero. In this stationary situation, past investors of funds presumably continue to collect a share of the income of the society, in the form of interest payments. Wages and rents are in balance with the rates of substitution between labor and land, and also balanced against the

interest rate at the margin of substitution of capital goods for other factors. If total investment is very large, the interest share of income may be quite substantial. The use of capital goods (as compared to the alternative of nonuse), however, has the effect of raising competitive wages and rents through increased efficiency or productivity, only a part of which goes as an interest reward to investors.[11]

There is one possible difficulty with the balance in this situation, however, with which economists of recent years have been concerned. If an economy approaches a stationary situation of the sort described, new investment opportunities vanish. There is then no place for the net money savings of the economy as a whole to go.[12] If now individuals still insist on saving out of their current net incomes, their savings will not be spent on capital goods (for which there is no additional demand) and as a consequence current expenditure will fall below current income. If this occurs, a progressive contraction of money income will be engendered, with resultant downward adjustment of money prices or employment or both, and will continue at least until unemployment sufficiently impoverishes the economy to reduce savings to zero. But underemployment would be a severe price to pay for stability.

This apparent dilemma is, of course, not inescapable. It could be escaped in the first place if people would simply stop saving altogether—the proper tactic in a stationary economy. Or it could be overcome if the situation were *not* stationary but sufficiently dynamic to create continual new net investment opportunities which would lead to net investment and get the savings spent as they were made. It is to the latter possibility that we ordinarily look. Let us therefore consider the relation of dynamic change in techniques, products, and tastes, and also in population and known resources, to the demand for capital goods and for investable funds.

[11] See Joseph A. Schumpeter, *The Theory of Economic Development* (Cambridge, Mass.: Harvard University Press, 1934), for the leading statement of the properties of a stationary general equilibrium of this sort.

[12] Excluding net demands for investable funds by the offer of old assets, etc., for sale—a net demand which will in any event drop to zero in stationary equilibrium.

[In the immediately succeeding discussion, the term "investment" refers to the use of capital goods in production and to the use of funds to finance their use; "total investment" to the total stock of capital goods accumulated to any one date, and to the total accumulation of funds tied up in financing them; "net investment" to the addition to capital goods used, to a corresponding addition to funds used, and thus to net spending on capital goods in the period of the net investment; and "net disinvestment" to a reduction in the amount of capital goods used, and thus to a failure to reinvest previously invested funds.]

DYNAMIC CHANGE AND NET INVESTMENT IN CAPITAL GOODS

Whereas in a situation which is stationary with respect to techniques, products, consumer tastes, and supplies and employment of labor and land, there may be a finite and limited demand for capital goods at each rate of interest, changes in any of these determinants of the economic equilibrium may alter the demand for capital goods; a steady progression of such changes may continually create a demand for net investment in such goods. With all these things given, as we have seen, firms in the economy would substitute capital goods for other factors down to a limit point, where, if stability of money income and employment could be preserved, there would be no net investment. A change in any of them, however, can lead to a change in the amount of capital goods employed and in the amount of funds invested.[13]

Let us consider the effects (1) of changes in the amounts of the primary factors, labor and land, with principal emphasis on *increases*, either by population growth or by discovery of new resources; (2) of changes in the techniques of production, generally improvements; and (3) of changes in the composition of the final group of consumer goods produced, either by the invention of new products, accompanied by a shift of consumer

[13] Further, changes in the money price level may cause changes in absolute money amounts invested; changes in the relation of capital-goods prices to their costs may influence real and money investment. These changes are neglected for the moment but will be referred to at a later point.

tastes in their direction, or by the shift of consumer tastes among known goods. In connection with the growth of supplies of factors, we will also necessarily consider the effects on investment of changes in the level of employment.

An increase in the supply of labor or of resources is generally accounted as an important force contributing to additions to the capital stock. Such increases have occurred historically because of population growth and of the discovery of new lands and resources; they may also occur if in a given population a larger proportion of persons offers themselves as labor (consider the shift of women away from the home) or if of given resources a larger amount are made available for use (as, for example, if government oil lands previously closed to drilling are opened). In either event, it is common sense that an increase in labor or resources *employed* will tend (in a given state of techniques, products, and tastes) to lead to some corresponding increase in the amount of capital goods employed. If we take two island populations of 100,000 and 1,000,000 persons respectively, with an identical ratio of resources to people in each case, and identical products, techniques, and tastes, it is quite evident that the larger economy will tend to use more capital goods (about 10 times more) than the smaller. If we contemplate the growth of one economy from the smaller to the larger size, in both people and resources, it seems almost inevitable that the capital stock would be added to as the growth occurred. The same would apply to growth in resources alone or labor alone. The discovery of valuable oil reserves in an island economy of given population and techniques should lead to investment for their exploitation; an addition of 25 percent to the population of such an economy should lead to a larger complement of capital goods with which the population can work. If ten shoemakers require ten hammers, twenty shoemakers require twenty hammers.

The analytical demonstration of these rather obvious facts is not difficult. In any given stationary situation, a general equilibrium is approached in which there is an optimum *ratio* of capital goods to the other factors—an optimum degree of routing of the services of these factors through the making of capital goods. This ratio is determined by the substitution relations affecting the use of capital goods with existing techniques and products,

and by the going rate of interest on funds. If the supply of either labor or resources is augmented, the economy should move toward a new general equilibrium. This movement, of course, requires either an upward adaptation of the flow of money income, to allow more to be employed at the same money factor prices, or a downward movement of money factor prices relative to money income, or both.

Difficulties may or may not be encountered in this process of adaptation, but unless the adaptive mechanisms are poor, the economy will tend to end up employing at least some of the additional resources and producing a greater output. If in this process of assimilation into employment of added labor or land, the capital stock were only maintained, the ratio of capital goods to other resources would be reduced, and the marginal rate of substitution of capital-goods services for other services should tend to be reduced. At the same time the price of capital-goods services should adjust in the long run with any changes in other factor prices, so that the ratio of the cost of machine hours, for example, to other factor prices should not be greatly affected. Then with a given interest rate on funds, the marginal rate of substitution of capital-goods services for other services would become less than the ratio of the prices of the others to that of capital-goods services, and firms would naturally substitute capital goods for other services (thus adding to the capital stock) until a new balance was struck. In a word, the tendency of firms always to minimize cost would lead them to maintain (with given techniques and products) a certain ratio of capital goods to land and labor; if the amounts of those factors in employment is increased, maintenance of such a ratio will involve an addition to capital goods.

From this it is apparent that growth of population or of known resources, as it occurs through time, may tend to create a demand for *net* investment in additional capital goods and thus a demand for new investable funds. A continued growth of population and resources might thus account for a continued demand, from year to year, for additional investable funds. The condition necessary for this to be true is that the economy should be able to assimilate at least some of the added resources into employment. If this can happen, the aggregate of employment,

output, and real demand for goods should be increased, and with it the demand for capital goods.

The place of population and resources growth in stimulating investment has nevertheless been the subject of controversy (1) because of doubts over the tendency of money income and money factor prices to coadjust to accommodate additional employment, and (2) because of doubt whether the added resources are supposed to lead to additional investment *before* or *after* they are employed. The second issue is of especial importance because the net investment corresponding to the new resources might be relied upon to create the additional spending which would be strategic in assimilating these added resources into employment. Historically, these doubts have not found extended support—that is, added population and resources have in the long run secured employment and the stock of capital goods has increased. But the historical process did not involve population and resource growth *in vacuo,* so that the record is not conclusive with regard to the logical points involved. Without becoming involved extensively in the theory of employment, we may comment briefly on the issues raised.

It is quite conceivable that money income and money factor prices might behave in relation to one another in such fashion that additional employment would not be accommodated by automatic economic processes, *unless* spending were augmented *first* by the purchase of additional capital goods—by net investment. In this event, added resources might be employed either if additional investment occurred "for other reasons" (*i.e.*, other than the growth of population and resources), or if the added resources led to net investment *in advance* of their employment.

Investment "for other reasons," of course, might or might not occur. If it did, and this resulted in employment of added resources, this added employment would further augment the demand for capital goods and for investable funds, and the "investment potential" of the new resources should be realized. But lacking this, would the simple presence of unemployed resources, added by growth, create a net investment demand prior to their employment? It seems obvious that the existence of resources which the economy shows no tendency to employ will not necessarily lead business firms to invest in anticipation of

employing them. It is not impossible that growth of population or resources might lead businessmen who believed, perhaps wrongly, that such resources *would* be assimilated in employment, to undertake net investment on the basis of this anticipation, and thus possibly to create the conditions for such assimilation. But it is also quite possible that, lacking these anticipations, added resources could simply remain in stagnant unemployment.

Summarizing the preceding, there would seem to be four alternative possibilities with respect to the effect of growth in population or known resources on the level of investment and the creation of additional demands for investable funds:

1. Money income and money factor prices automatically coadjust so as to assimilate the added resources without any necessary prior assistance from net investment—the economy thus automatically always moves to full employment. In this event, added resources automatically result in proportionate additions to total investment in capital goods, and progressive growth is a certain source of a recurring net demand for investable funds.
2. *Instead*, money income and money factor prices do not coadjust automatically to assimilate added employment, and the added resources will lie idle unless there is first an addition to spending via the purchase with investable funds of new capital goods—*i.e.*, a spontaneous rise in spending. If this is the situation, we have three subcases: (*a*) Added investment fortuitously occurs for other reasons. This draws the added resources into employment, and as they are employed, their employment results in further net investment. In this case, added resources do result in further investment *per se*, but this has to be *induced* by other net investment, which must therefore play the leading role in the process. (*b*) Businessmen believe that the added resources will be employed, and therefore undertake added investments in anticipation of an expanding market. This investment leads the added resources into employment, with further additions to investment, and the businessmen therefore turn out to be "correct" in their

anticipations. This is a possible pattern, but relies on the "self-correcting error," or "self-justifying optimism," of investors. (*c*) Businessmen do not make this "error," but wait for the added resources to be employed. The resources are not, and remain unemployed, with no additions to investment, until some fortuitous circumstance of the sort cited in (*a*) above occurs.

Whether, in general, actual experience conforms to the first pattern or to one of the three versions of the second is a matter of some disagreement among economists. We will discuss this matter further in a later part of this chapter. In general it will appear, however, (1) that added resources have a good chance of generating additional investment in capital goods and additional use of investable funds; (2) that they may not automatically do so; and (3) that if their stimulus to investment is not automatic, then its realization is likely to depend upon other sources of net investment demand or upon a fortuitously optimistic attitude on the part of investors.

It should also be emphasized that if the economy does not automatically cling to full employment, the assimilation of added resources is not necessarily permanent. And just as the addition of resources to employment may lead to additions to investment, a subsequent decline in employment may lead to subtractions from investment, or disinvestment. Thus the "new equilibrium" of investment corresponding to the consummated assimilation of new resources may in fact be temporary in character. Growth in the amounts of labor and land available may thus in general lead, for one reason or another, to net investment over time. But the specific level of investment demands is simultaneously conditioned by movements in the general level of employment.

This is consistent with the principle that the total of capital goods used is related to total employment. With any supply of labor and land, the amount of capital goods used will depend upon the proportion of the resource which finds employment. One level of investment will correspond to full employment, and a lower level to 60- or 80-percent employment. If there are barriers in the way of stable full employment, a part of available

resources may remain unassimilated and total investment—*i.e.*, the value of the total accumulation of capital goods—may be correspondingly restricted.

Movements for any reason in the level of employment of a given supply of resources will lead to corresponding variations in the level of capital goods used. A rising employment will generate net investment; a declining employment will cause net disinvestment. The mere existence of unemployed resources, which are not being assimilated into active use, will not necessarily cause the net investment activity which might employ them. But those forces (including other sources of variation in the rate of net investment) which cause the level of employment to increase or decrease, will thus be able to *induce* net investment or disinvestment. Any force which causes the level of employment to rise will generate *induced investment*, in the form of additions to investment to match the rise in employment, and any force which causes the level of employment to decline will generate *induced disinvestment*. It is also worth noting that with rising employment the induced net investment occurs because of the *increase;* and that the added investment demand is quickly satiated as soon as employment is stabilized at a higher level. The induced net investment (or disinvestment) demand is the product of *change* in employment and drops to zero as soon as the change is completed.

The employment-investment relation is seen most strikingly in the periodic fluctuations of employment which characterize the business cycle. The cyclical upswing involves an increase in the flow of money income which induces an increase in employment; this creates an additional demand for capital goods and for investable funds. In the corresponding decline in money income and employment, the demand for capital goods declines, and some disinvestment will take place, releasing previously invested funds. The fluctuation in the use of investable funds will result not only from net investment and disinvestment in relatively short-lived capital goods, but also from variations in the rate of reinvestment in more durable capital goods. Thus in periods of declining employment and output, reinvestment for replacement of such goods will tend to be postponed, and in periods of rising employment reinvestment will tend to be hur-

ried. A fluctuating money income and employment thus tends to create a directly correlated fluctuating flow of net investment and reinvestment of funds. Since such investment tends to augment expenditure, and lack of it or disinvestment tends to reduce expenditure, this fluctuation of investment spending accentuates greatly any initial tendency to fluctuation of money income which may exist in the economy.[14]

TECHNOLOGICAL PROGRESS AND NET INVESTMENT

Growth in the amount of employed labor and natural resources is, of course, not the only sort of change which may create a demand for additional capital goods and investable funds. At least equally important historically are the development of new techniques of production and the introduction or rise to importance of new products. Both sorts of change have a tendency to lead to some net investment.

The general relation of techniques to the demand for capital goods is quite clear. With any given set of known techniques, there is a corresponding set of substitution relations between capital-goods services and other productive services. It is in terms of a given set of techniques that the family of isoquants showing these relations among factors is defined. This is sometimes expressed by saying that for any state of techniques there is for each industry a corresponding *production function—i.e.*, functional relation between output and relative and absolute amounts of various factors employed, as expressed in a family of isoquants. For the economy, there is a given set of such production functions, corresponding to given techniques. With such production functions (and with other determinants unchanging) the economy will arrive at an equilibrium use of capital goods corresponding to any attained level of employment and total output. When new techniques are introduced by business firms, these innovations take place generally because they allow the innovators to reduce costs of production, and thus, transitionally

[14] See A. H. Hansen, "Economic Progress and Declining Population Growth," *Readings in Business Cycle Theory*, G. Haberler, ed. (Philadelphia, The Blakiston Company, 1944), Chap. 18, for a discussion of one thesis concerning the impact of population changes.

at least, to make an extra margin of profit. (The emergence of pure profits from technological change will be discussed in Chapter 14.) This means in turn that efficiency is increased, or, in general, that less total resources are required to produce a given unit of output. A reduction in cost per unit of output, however, is not the sole effect of innovation; in fact, several effects may be distinguished.

First, the innovation will alter the production function in the field affected, to the end (1) that different proportions of the various factors, including capital goods, may be profitably employed to produce a unit of output, and (2) that different absolute amounts of the various factors may be employed. This may mean in effect that either more, less, or the same amount of capital goods or of investable funds will be employed to produce what was the preceding equilibrium output. Second, the cost of production of the affected output will be reduced, to the end that, as the new equilibrium is sought, a larger output will tend to be produced, provided that total employment has not been reduced by the process of change. So far, therefore, as over-all employment tends to be sustained during the process of change, the increase of output because of increased efficiency tends to lead to added investment. This may augment any increase in investment realized at the previous output, or offset in a degree any reduction realized at the previous output. Third, the combination of output extension and increase of efficiency will either increase, hold constant, or reduce the quantity of labor and land employed there. Except where the employment in the affected industry holds constant, there will thus be possible readjustments in other industries which may affect investment, though not in any certain fashion.

Supposing that we sum the three effects so far mentioned, and assume that total employment in the economy is maintained, what will be the net effect on total investment of a technological change? It might conceivably be negative rather than positive—that is, the investment (at a given price level) per unit of labor or land might conceivably decrease. For this to be true, however, the saving in the use of capital per unit of output in the affected field would have to be great, and the investment added in all fields in the process of assimilating released resources

would have to be small. Since there is no reason that the innovation must tend to economize in the use of *capital,* it is quite possible that total investment will be increased.

A fourth effect may be of as great importance as any other. If the technological change takes place in an industry producing *a capital good,* and thus reduces the cost and price of this good and stimulates its purchases, then if that good has a greater than unit elastic demand, total money investment by buyers will increase as the result of the technological change. Taking all the forces into account, it seems rather likely that technological changes will tend to increase total investment. But such a positive effect is not logically certain.

The general impression that the effect will be positive is derived largely from historical experience. The "industrial revolution," which has been essentially one long progression of major and minor innovations of technique, definitely did tend to increase both the ratio of capital goods to other factors and the total amount of investment. Great and small innovations such as the steam engine, the railroad, the coke blast furnace, the continuous rolling mill, the central generation of electric power, machinery for stamping out metal parts, etc., have led to an increased use of capital goods and to a steadily increasing demand for investable funds. Many of the major innovations involved the invention of a new sort of capital good to replace a non-capital-using method employed before. Very often, therefore, the investment per unit of output, as well as investment per unit of employment, was definitely increased as the result of innovation. This was apparently true of the case where the Bessemer converter replaced the earlier steel furnace, or where the railroad replaced the horse-drawn carriage for inland transportation. In other cases, of course, such as the replacement of steam-drawn railroads by motor buses for suburban commuters, the investment per unit of service was probably reduced, and perhaps also total investment.

But in citing logical possibilities and isolated instances from history, we perhaps take too narrow a view. Changes in production technique are revisions of method which result in reducing costs or ultimately in using less labor and land (direct and indirect) per unit of output. The obvious way of doing this, if we

read our economic history, is to find more and better ways of using capital goods—of investing our primary resources initially in machines and tools for production. It may thus be that almost by nature the process of technical innovation is a process of accumulating capital goods and adding to investment. If there are episodes where an innovation happens to reduce the amount of total investment, they probably depart from the general tendency. When the resultant growth in efficiency and output is allowed for, technical innovation seems inevitably associated with increasing total investment of funds and use of capital goods.

We conclude that, although technological innovations may conceivably have the effect of reducing total investment, it is highly probable that on the average they will lead to added investment of funds in capital goods, and will thus be a source, as they recur, of a re-emerging net investment demand. This is contingent, however, on the assumption that they do not result in a reduction of over-all employment, with a consequent offset to any increase in total investment. Is it likely that the potential investment-increasing effects of technical change will be negated by corresponding reductions of employment? This is an extremely broad and difficult question, proper treatment of which involves consideration of the responsiveness of commodity prices to cost changes, of the mobility of labor and land among occupations, and of the precise *sequence* of effects following a technical change. We are not prepared at this point to examine it thoroughly.

One general comment, however, may be pertinent. A significant accumulation of technological change, by increasing efficiency, tends to increase the total output and real income corresponding to a given level of employment. This may, in turn, induce individuals of inflexible habit to consume a smaller proportion of their incomes, and thus to save more, which in turn can be spent only via net investment. There may conceivably be some tendency for consumption spending, as a proportion of income, to fall off, at least temporarily, in response to increased efficiency, and for employment to be virtually reduced—in effect, an unwillingness of the economy to assimilate the additional resources "freed" by increases in efficiency. Aggregate consumption

demand in real terms, that is, may not be immediately flexible. If this occurs, technological change could reduce employment and thus have reduced or even negative effects on total investment. But historical observation argues strongly that the net effect of the sequence of technological changes has been to add regularly to the total amount of investment, and not to lead to a *progressive* accumulation of unemployment. Thus we will perhaps make no great error if for general purposes we refer to technological changes as sources of recurrent net demands for investable funds.

As a potential source of additional demand for capital goods and investable funds, we must also include changes in products and in buyers' choices among products. Simple shifts in taste among a given list of products already in general use may, of course, have some effect on investment demand—increasing the use of capital goods in the production of the good with growing demand and restricting their use in the production of the good that loses out. The ultimate net effect on the investment of funds in this event might be positive, negative, or neutral, depending upon the relative importance of capital goods in the two lines. Thus a shift of consumer demand away from brick and stone houses and in the direction of frame houses might increase investment in lumbering, sawmills, and nail making but restrict investment in brick plants, quarries, and the manufacture of mortar, and the *net* effect on investment might be of any sign. It is probable, however, that the *initial* effect of any marked shift in demands will be to increase net investment of funds, since new investment in the expanding field will probably take place more rapidly than disinvestment in the declining field.

The prominent instances of such shifts in demand, however, do not involve two old and well-established products but rather the introduction of a new product which provides a new service, or a known service in a much improved form. Thus the histories of the radio, the automobile, or the electric refrigerator involve a substantial innovation in product followed by a development of consumer demands in favor of the new product, a substantial addition to total investment in capital goods, and a corresponding addition to the total real output of the economy. In effect

a process of addition rather than one of substitution was involved. May such innovations be depended upon to create new net investment opportunities recurrently through time?

It is not apparent that they may be depended upon to do so *per se*. The availability of autos and radios to the poverty-stricken masses of India or China has not led to any substantial amount of investment for the production of these goods to supply those markets. In those countries, all productive effort is required to provide food, shelter, and clothing, and a shift of demand in favor of the automobile would be fanciful.

It would appear that the innovation of "new products"—providing new services or old ones in improved form and in much greater quantities—will be possible on a large scale only where progress in general efficiency has so cheapened other goods that their consumption can be maintained while adding the production of new goods. If technological progress, or abundance of natural resources, can "free" sufficient resources from the production of the going bill of goods—that is, if real incomes of consumers are high enough—then demands *may* emerge for newly developed products which will employ the resources "freed" from other lines. Or, in essence, the total real demand for consumption goods may most effectively be increased, in an economy which becomes richer and richer, by the addition of new goods for consumption. This will in turn lead to additional investment for the production of the new goods, and create a net demand for investable funds. But the strategic conditions for new products to create such investment demands would seem to be (1) progress in other lines—in techniques of production or in the amount of natural resources available relative to population—which progressively enriches the economy so that it can add to its bill of consumption goods and (2) sufficient *additional* demands for the new goods.

Innovations of product may indeed be strategic in generating the demands which keep the resources of a progressive economy fully employed. If, of course, the new products also increase efficiency in production, their effect is multiplied. We may conclude that (1) simple shifts in demand among known products will in the long run tend to have a neutral average effect on investment, and (2) that introduction of new products may help

create net investment demand so far as the productivity of the economy (per capita) is increasing, or can be increased by employing idle resources, and so far as the new products attract an addition to consumer demand for the economy as a whole.

The positive independent contribution of new products is, then, as follows. In a progressive economy with its resources at any rate rather fully employed, the new product may in a sense simply take the place of expansion in old products, and either would require added investment. It may be most important in averting any tendency for expenditure of income on consumer goods to drop off and thus to precipitate declining money income and employment—in preventing successive improvements in efficiency from creating unemployment. If an economy is already stagnant with unemployed resources, because money income and money factor prices will not coadjust to permit full employment, the introduction of new products may increase consumer spending out of the incomes of employed resources, and thus lead to net investment and added employment for producing the new product. This effect on investment, however, must depend either upon the ability of new products to increase the ratio of consumer spending to income, or upon the inclination of producers to make investments in the anticipation that this will happen.[15]

THE TIMING OF NET INVESTMENT

Let us recapitulate our observations on the source and nature of the demand for capital goods and for investable funds. In any given state of techniques, products, consumer tastes, supplies of labor and land, and employment, there tends to be a finite and satiable demand for the services of capital goods, governed by the prices of such goods (in ratio to their costs in wages and rents), by the rate of interest, and by the substitution relations between capital-goods services and other productive services. In such a given stationary state, therefore, a given stock of capital goods will tend to be acquired and, once acquired, regularly re-

[15] For further discussion of technological change and investment, see A. H. Hansen, *Fiscal Policy and Business Cycles*, Chaps. 1 and 2; and William Fellner, "The Technological Argument of the Stagnation Thesis," *Quarterly Journal of Economics*, vol. 55, pp. 638-651.

placed to maintain the stock at a given level. Correspondingly there will be (with any given money price level) a finite demand for funds to be invested in these goods at the going rate of interest. Once this investment has taken place, additional investment of funds will not be required; the net demand for new investable funds will drop to zero; and the demand for reinvestment of funds will hold constant.

Progress or change in any of the determinants of investment, however—in techniques, products, tastes, or supplies and employment of basic resources—may change the demand for capital-goods services and create recurring net demands for investable funds. Growth of population and resources, improvement in techniques, introduction of new products, and changes in consumer tastes all have some potential effect on investment, and in general they all (with the exception of simple shifts in taste) tend to increase it and to create additional demand for investable funds. This has been true historically, and logic indicates that such a result is probable if not strictly necessary. In the case of additions to population and resources, it is obvious that such resources require something like proportionate additions to capital with which to produce. In that of technical innovations, the *nature* of such innovations has been to reduce costs and increase efficiency by finding more and better ways of using capital goods to produce with, and further to increase the total output potential so that the total need for capital goods was progressively augmented. As regards new products, they have served in general to take up the slack created by increasing general efficiency and to provide demands for investment to produce additional output. Combined progress in techniques and products seems to be inextricably associated with recurrent net investment in capital goods.

In the preceding discussion of net investment demand, however, an issue keeps reappearing concerning the *timing* of net investment activity in response to basic changes of the sort enumerated. When population grows, or techniques change, for example, will net investment demand be forthcoming *at once* in full measure or will it wait upon the approach to a new employment equilibrium appropriate to the changed situation? Now of course if the economy always, after every change,

moved without difficulty to a full-employment position, this would be an idle issue—the full investment potential of every change would always be realized. The problem is made an issue because of the belief of many economists that the economy may *not* easily adapt to a condition of full employment—because of a shortage of money income relative to the level of money factor prices—and may not easily assimilate the additional resources supplied, for example, by population growth or previously freed by the economies of technical innovation. This is combined with the hypothesis that added *net investment*, or demand for investable funds, will serve to augment money income and employment. The issue is therefore posed: If we have given a stagnant money income and employment, which will not automatically adjust to admit more employment, in what degree will changes in techniques, population, and so forth, create net investment demands *in advance of* an increase in employment? This issue has been discussed at various points in the foregoing pages, but it may be well to recapitulate our conclusions here.

A precise and detailed answer must, of course, depend upon a correspondingly precise description of the character of the stagnation difficulty which the economy is supposed to face, and we are not disposed here to explore this problem at length. A simple version, however, would involve the assumption of arbitrarily *rigid* money factor prices for labor, land, and (therefore) capital goods, and a money income which is stagnant at a level insufficient to produce full employment and augmentable only in the event of additional spending on capital goods financed by new investable funds.[16] In such a situation, how far may we count on various aspects of dynamic change to generate the additional net investment which will generate more income and employment?

Under the assumptions drawn, growth in population or in known natural resources could not generally be counted upon to create net investment demands *in advance of employment of the added resources*. Such growth would simply augment any existing unemployment without creating additional effective mone-

[16] Reductions in money factor prices are ruled out either as impossible or as self-negating because always matched by corresponding changes in money income.

tary demand, the limitation of which is the basic trouble. A spontaneous increase in spending, however, from other causes, might draw such resources into employment, at which time their use would require added investment.

Changes in the techniques of production are potentially reliable sources of added investment demand even in a stagnant situation of the sort described. That is, new capital goods will be installed to meet the going levels of demand, and these installations will usually require the use of additional investable funds. If they do, a stimulus to money income should result, which may in turn assimilate added resources and take full advantage of the increased efficiency. If they do not, of course, the effect is limited to a simple shift to more efficient techniques for supplying a limited demand.

Shifts in consumer tastes among products may also create a requirement for net investment in a stagnant situation, particularly so far as net investment in the growing lines proceeds more rapidly than disinvestment in those which lose out. This argument applies also in general to newly introduced products to which consumer fancies turn. The simple introduction of new products, however, which potentially add to the consumer's bill of goods rather than replacing other items, may not be counted on for spontaneous increase in investment and employment unless (1) the economy is "rich enough" to support production of the added good, and (2) the availability of the new product creates the actuality or the effective prospect of a spontaneous increase in spending to acquire it.

One implication of the preceding for the general theory of investment and the use of capital deserves mention. Under conceivable circumstances the economy may hope for the continuation of dynamic changes, which potentially provide recurrent net demands for investable funds and thus keep money income and employment at a high level. But some potential sources of investment demand are not necessarily actual sources, and the effects of dynamic change may accumulate in a latent form as unemployment remains or increases. If continual adjustments to full employment are not made, then the pent-up but latent investment demands may still be released and made effective by any expansion of money income and employment. Especial im-

portance may therefore attach to recurrence of those dynamic changes which may set off or initiate expansions of investment, money income, and employment time after time. In this connection, it would appear that technological changes together with introduction of new products will be of primary importance in maintaining a high level of net investment activity and employment. We shall return to this matter later in this chapter.

NET INVESTMENT AND GROSS INVESTMENT

In the preceding pages we have discussed the determinants of investment in capital goods and of the rate of additions to this investment over time. Analysis indicates that in any stationary situation, or in any currently given situation which was maintained in the total absence of further dynamic change, there would be a finite total demand for capital goods at any rate of interest, and a determinate schedule of total demands for capital goods at various rates of interest. Corresponding to this demand schedule for capital goods, there would be at any given price level a determinate total demand schedule or curve for dollars of investable funds for use in financing capital goods. This schedule would show the total amounts of investable funds which producing firms as a group would wish to have invested in capital goods at various possible rates of interest. As dynamic changes occur, and new situations are introduced, there may be additions to the demand for capital goods and for investable funds, so that the total demand schedule for investable funds effectively shifts to the right. With steady progress in an economy, we might expect on the average that in each successive year the total demand curve for investable funds would lie farther to the right.

This may be illustrated by a succession of total demand curves for funds referring to successive points in time, separated for example by intervals of a year, as in Figure 52. Here D_1 represents the total demand schedule for funds at the beginning of some year 1, D_2 the total demand at the beginning of the following year 2, D_3 the total demand at the beginning of the year 3. The rightward shifts might be supposed to occur because of technological changes which generate additional demands for

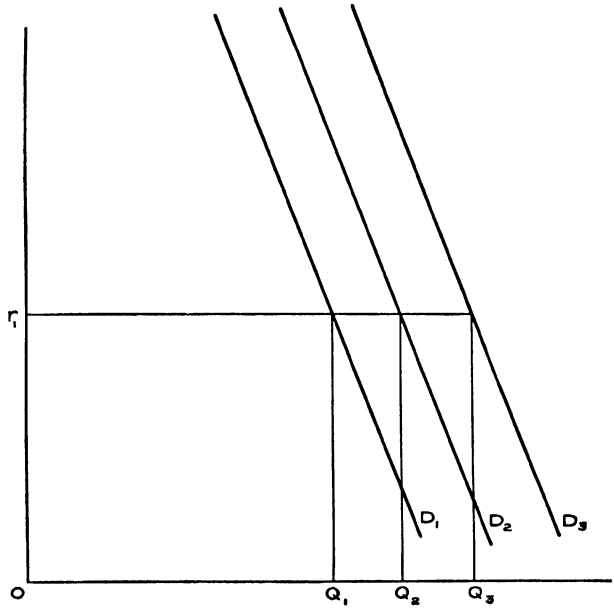

Figure 52

capital. The implications of this sequence of schedules is generally as follows: If the rate of interest is at some level r_1, the economy will tend at the beginning of the year 1 to have invested the quantity of funds OQ_1; at the beginning of year 2 it will have tended to bring total investment to OQ_2; at the beginning of year 3 it will have approached the total investment OQ_3. Similarly, the addition to investment of funds *during* year 1 will tend to be Q_1Q_2, and during year 2, Q_2Q_3. These additions represent *net investments* in capital goods—increases in the total amount of funds invested in capital goods in successive periods.

This last item of information is quite significant because it brings us close to what matters most for purposes of determining the rate of interest—namely, the added demand for funds occurring in particular intervals of time, or the current drain on the flow of money payments resulting from net investment activity. It is perhaps interesting, that is, that in all previous time to date at the beginning of year 2, the total investment of

funds OQ_2 will have occurred. But for analysis of the interest rate during a current year 1 or 2, this is not so significant as the *current* demand for funds. It may therefore be well to investigate the technical meaning of the concept "an additional demand for funds," such as Q_1Q_2, during year 1.

This is at all complicated only because, during any current period, funds *previously* invested in capital goods (in past periods) will be recovered from operations, and will require *reinvestment* if the going level of investment is to be maintained. Up to the beginning of year 1, let us say, the amount OQ_1 of funds has been invested in capital goods of all sorts—let us suppose OQ_1 represents 200 billion dollars. As time passes hereafter, however, the 200 billion dollars' worth of capital goods will be used up or wear out and, as this happens, the funds invested in them will be recovered from sales receipts by enterprise. To maintain the initial level of investment of the beginning of year 1, these funds will have to be reinvested, and a *gross* rate of current investment sufficient to maintain the level *will be counted as meaning that there is zero net investment.* The rate of gross investment (made in a time interval) necessary to maintain the total level of investment effective at the beginning of that period may be designated the *equilibrium rate of reinvestment.* What this rate will be for 200 billion dollars will depend primarily on the average *durability* of the capital goods in which past investment has been made—that is, on the rate at which they wear out or are used up. Thus if the *average* durability of total investment is 10 years, the equilibrium annual rate of replacement (assuming a constant rate of use) will be 20 billion dollars. If we take this as a reference point, then, the meaning of an additional or net investment during year 1 of Q_1Q_2 (say 15 billion dollars) is that there is this amount of current investment of funds *in addition to* 20 billion spent on replacement, or that *gross investment* made in the period is 35 billion, which exceeds equilibrium reinvestment by 15 billion. Net investment thus refers to funds currently invested above "normal replacement" of the capital goods on hand at the beginning of the period. The emergence of net investment in any period is evidence of adjustment to dynamic changes, the investment-generating effects of which have not previously been

exploited. It will be noted that the rate of net investment per year need not (even at a given interest rate) remain constant from year to year. The rate of dynamic change is freely variable, so that net investment may grow, decline, or fluctuate as time passes.

[In the discussion hereafter the term "*gross investment*" will be used without modifying terms to refer to "gross investment spending *of a current period*"—*i.e.*, to amounts spent in such a period on capital goods output. The components of such gross investment are *reinvestment*—funds currently spent to replace capital goods; and *net investment*—funds currently spent to add to capital goods. The "total (level of) investment" still refers to the accumulation of all *past* investments in capital goods as of the beginning of the current period.]

So far we have referred to the addition to investment per period at a given interest rate. This may be extended to a schedule of net investment demands in any period at various possible interest rates. Such a net investment demand schedule would show for a given year (or other time period) the net additions to or subtractions from an equilibrium rate of reinvestment for that period. It would in general be derived from Figure 52 above for the year 1, for example, by taking as net investment quantities at various interest rates the distances of the schedule D_2 from the solid line extended vertically upward from Q_1. Or, precisely, we would show, as in Figure 53, (1) a schedule in the form of a vertical line, q_1q_1', representing the equilibrium rate of reinvestment (to maintain investment at OQ_1) for the year 1, and (2) a net investment demand schedule, d_n, showing additions to or subtractions from this rate of reinvestment at various rates of interest. The amount oq_1 corresponds to the 20 billion dollars of reinvestment necessary to maintain investment at OQ_1 (or 200 billion) through the year 1, and the schedule d_n shows the net investments (or, above the rate r_a, net disinvestments) to be added to or subtracted from oq_1 to get the gross demand for investable funds during year 1.

It will be noted that the net investment schedule d_n does not refer strictly to added investment in *new capital goods*, but rather to such investment plus or minus any variation in reinvestment from the equilibrium rate, resulting from interest rate

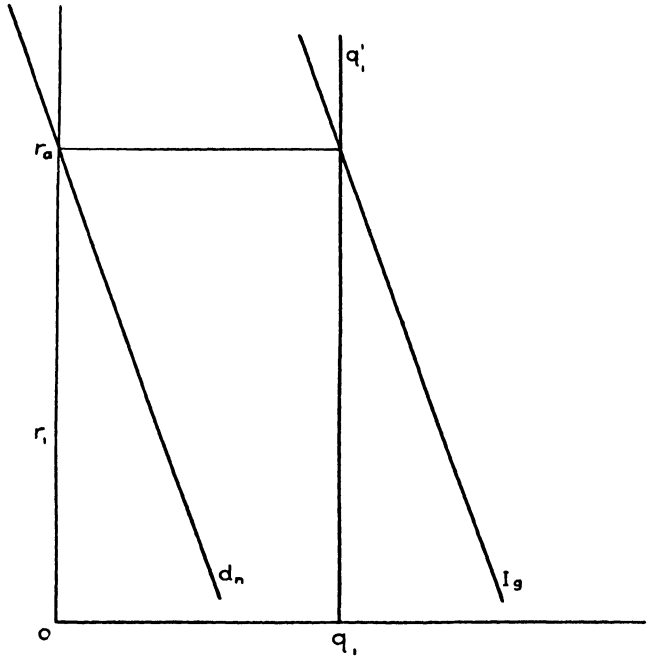

Figure 53

changes or other causes. Thus (referring to Figure 52), suppose that investment at the beginning of year 1 has been carried to OQ_1 at the rate of interest r_1, with the equilibrium rate of reinvestment corresponding (Figure 53) at oq_1. Now the backward slope of the net investment schedule d_n above the interest rate r_1 reflects not only the decline of *new* investment in response to rising interest rates but also the restriction of reinvestment below oq_1 because of the same interest rate change. The *net investment* demand schedule, as departures from an equilibrium rate of reinvestment (or from "capital maintenance") thus shows variations in new investment plus variations in the rate of reinvestment in response to changes in the rate of interest.

The two schedules d_n and q_1q_1' (net investment plus reinvestment as defined) are added together to constitute a *gross investment* demand schedule for the year, Ig, in Figure 53; this schedule shows the total demand for investable funds for re-

placements plus additions to capital goods for a given period. This gross schedule, which may hereafter occupy our main attention, shows that at the rate of interest, r_a, where it intersects the line q_1q_1', gross investment of the period equals the equilibrium rate of reinvestment; below this rate there are various amounts of net investment, and above, various amounts of net disinvestment. Some such gross investment demand schedule will rule in each successive period of time, *showing the amount of funds demanded for purchase of capital-goods output at various rates of interest for that interval.* A positive rate of dynamic change from period to period will mean that in any period a considerable range of this schedule will be to the right of the equilibrium rate of reinvestment (oq_1), thus indicating an additional demand for funds during the period. A negative rate of dynamic change (represented perhaps in declining employment) will mean that a considerable range (and perhaps all, down to a zero rate of interest) of the schedule, will lie to the left of the equilibrium rate of reinvestment, indicating net disinvestment of funds for this period. The position of the gross investment demand schedule will shift over time in response to varying rates and directions of dynamic change.

With successive increments of net investment, the rate of reinvestment will become larger over time (with net disinvestments it will become smaller), and this changing rate of reinvestment will in any period be augmented or reduced by the net investment or disinvestment of that period. Gross investment also fluctuates if the *rate of reinvestment* is for any reason *irregular* over time, as it might be, for example, if very many replacements of durable goods came due all at once every five or six years, with a hiatus in replacements in between. If, however, reinvestment tends to occur at a regular rate (as it would if the same proportion of total investment wore out or was used up each year) then fluctuations in the position of gross investment demand schedule from period to period would reflect primarily the impact of a varying rate of dynamic change on the rate of net investment. As a progressive economy proceeded through time, with investment opportunity in each succeeding period, we would find (1) that the equilibrium rate of reinvestment would become progressively larger, as the total of invest-

ment grew, thus causing a general rightward shift in the gross schedule, and (2) that the net investment demand of successive periods would vary, thus causing additional movement or fluctuation in the gross schedule.[17]

As we pass from period to period, the net investment demand of the preceding period tends to have been fulfilled, and is now reflected in a different total investment and a different equilibrium rate of reinvestment. It has been reflected in net investment only once. Each successive period must look for its own net investment demand, from additional dynamic change. Each successive net investment demand becomes satiated, and new ones must continually emerge if gross investment is systematically to differ from a plain reinvestment rate.

Particular interest attaches to this net investment demand, or difference between gross investment demand and the current equilibrium rate of reinvestment. This is because on the average the equilibrium reinvestment rate makes no draft for funds on the net income of the economy, but is just matched by a share of gross income which is set aside as a part of the costs of production before payment or calculation of net income. Correspondingly, it is only net investment or disinvestment which makes a draft upon the net income of the economy. Suppose an economy has an investment of 200 billion dollars in capital goods of which 20 billion dollars comes due for replacement each period. Now if the economy is currently in equilibrium to this rate of investment, its income statement for a period might look as follows:

Total sales receipts of enterprises (assumed)	150 billion
Net income payments (wages, interest, rents, profits)	130 billion
Depreciation and replacement of total investment	20 billion
Total	150 billion

[17] It should be noted that both the gross and net investment demands for funds as defined will be reflected in exactly corresponding demands for current output of capital goods in the period of reference. The total demand for investable funds is defined for the economy as a whole, and thus represents an algebraic sum of the investments (positive and negative) of individual firms. Thus the disinvestment of firms which sell capital goods on hand from preceding periods, without currently replacing them, and thus take money out of the income stream is offset against the capital-goods acquisitions of all firms, and the algebraic sum represents a demand for current output.

In effect, the enterprises of the economy are receiving, as the aggregate demand price of their output, or *gross income*, 150 billion. They pay out of this 130 billion in net income, but have to retain, as a part of cost, 20 billion to finance replacement.[18] This is available for expenditure, but it does not come out of the net income of the individuals of the economy. It is "automatically saved" out of the payments, at equilibrium prices, for the commodity output of the economy, and is deductible in arriving at net income. Now if gross investment only equals equilibrium reinvestment, the recipients of net income have their entire incomes to spend (or not) on other than investment goods. Moreover, gross investment will cause the expenditure on investment goods of exactly all of the gross income which is not payable as net income. Net investment (or disinvestment) demand, on the other hand, implies an expenditure on investment goods at a different rate. Thus if gross investment runs ahead of the equilibrium rate of reinvestment, the difference (net investment) is a demand for funds in addition to those available after payment of net income, and this opens an avenue for the expenditure of the net income of the economy or of other supplies of funds. If gross investment runs below the equilibrium rate of reinvestment, the difference (net disinvestment) is a *subtraction* from reinvestment, meaning that an equivalent amount of funds normally ready for reinvestment are not so spent and thus virtually reduce the income stream of the following period. In effect, therefore, there is considerable significance attached to the current rate of gross investment in capital goods, as derived from the gross investment demand schedule, and in the relation of this gross investment to the supply of funds normally provided for reinvestment from the sales receipts of business enterprises.

As we turn to the analysis of the rate of interest we will recognize: (1) that in a stationary state the gross investment demand would occupy such a position from period to period that at a steadily maintained interest rate the average gross investment per period would eventually just equal equilibrium reinvestment, and thus be exactly financed from reinvestment re-

[18] And 20 billion in gross output goes to offset the wearing out or using up of capital goods.

serves of enterprise; and (2) that in a dynamic economy, the gross investment demand schedule will lie ahead of or behind this position—but generally ahead in response to growth and technological change—thus leading to net investment of additional funds, or disinvestment of previously invested funds, from period to period, according to the conditions prevailing in successive periods.

The preceding gives us, in a rather simplified fashion and with the omission of many details and minor modifications, a picture of the behavior of the demand for funds for investment in capital goods from period to period through time. Looking ahead from the present analysis to its application, we see that a significant aspect of this gross investment demand is that it is a demand for funds *which will be spent* so as to elicit current output and which will thus move through the system to create money income and elicit production and employment.[19] The amount *invested* or reinvested each year is an amount secured for spending on capital-goods output and so spent. Gross investment represents that share of gross income which is spent on capital-goods output during this period and which thus generates income for the next. In particular, it is that share of income which is spent *in addition to consumption.*

INVESTMENT IN CONSUMER FINANCE

This investment demand for capital goods is not the only demand for investable funds for expenditure. One potential addi-

[19] As such, it is similar to consumer-goods demand, which also elicits current output and creates income (except so far as offset by disinvestment), but different from the demand for idle cash balances to be held, which might be made effective by offering for sale land, securities, or old capital assets on hand because of the production of previous period. Such demands do not elicit current output. So far as capital assets available from earlier periods are sold to "investors" for current use without being replaced, and thus serve in lieu of current production, they do not result in an addition to investment (the disinvestment of the seller canceling out the apparent investment of the buyer) but rather are simply involved in an exchange which results ultimately in hoarding. This follows from the fact that current investment for the economy must be defined as the algebraic sum of investments and disinvestments by all investing units.

tion, or distinguishable segment of current investment demand, deserves especial mention—namely, the *consumer finance* demand. In the modern economy, investable funds are sought not only by business enterprises which wish to invest in capital goods for production of further output for sale, but also by individuals who wish to borrow to augment their current purchases of consumer goods. Such loans are sought principally to finance the purchase of expensive durable goods but are also sought to finance the acquisition of a wide range of items more frequently purchased. The demand for such loans expresses itself as a schedule of amounts sought, per period of time, at various rates of interest. The psychological origin of such a demand is presumably a preference by the borrowers for present goods (real income) as compared to future goods. They are willing to pay a rate of interest, thus somewhat reducing their eventual total spending power, in order to have money to spend now, when they borrow, instead of later, when they repay. The demand for consumer loans is thought not to be very sensitive to the rate of interest, but there is nevertheless presumably some interest elasticity, so that the demand for funds for consumer finance, I_c, might appear in any period as in Figure 54. This would show that, in the period of reference, certain amounts of dollars for consumer loans would be demanded at various rates of interest.

This demand schedule, like that for investment in capital goods, however, is also subject to satiability, so that reference is again necessary to gross and net demands. "Starting from scratch," as it were, a new demand for consumer finance would create a net requirement for investable funds. But if the initial borrowers are supplied with funds, in succeeding periods their repayments will furnish funds for new loans. It follows that if, from year to year, for example, people as a whole always demanded new consumer loans of 5 billion dollars, a year would soon be reached where repayments exactly balanced new loans, and where there would be no net additional demand for funds. Gross investment would just equal reinvestment at this point. Let us, therefore, understand the schedule I_c as a gross investment demand schedule for investable funds in any given period. The position of this schedule will be dependent primarily on

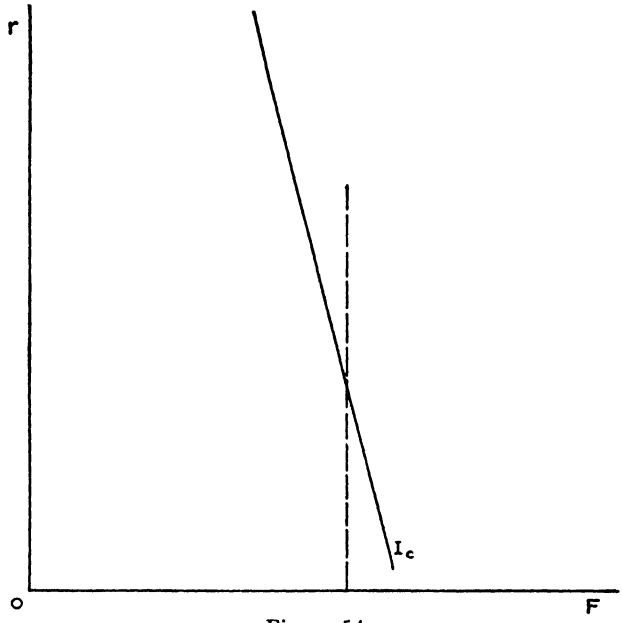

Figure 54

the level of income—shifting right with increasing income and left with declining income.

Corresponding to this will be an equilibrium rate of reinvestment, represented by the dotted line in Figure 54, which is in effect the rate of repayments of past loans in the current period. In any stationary situation, where income remains constant from period to period, the gross consumer loan investment at any going rate of interest will quickly become equal to the reinvestment or repayment rate, and no net investment demand will occur. Thus, in stable income and interest situations, we have a group of firms—let us call them finance companies—with a given supply of funds in investment which regularly rotates from one consumer loan to another, and there tends to be no investment or disinvestment. With movements of income, however, there will be shifts of the gross investment schedule, I_c, away from equilibrium, resulting, while income is on the move, in net investment or disinvestment. Net investment will mean that additional funds, above those previously invested and now

flowing back as repayments, will be demanded to finance consumer purchases. Net disinvestment will mean that repayments are running ahead of current loans, so that there are net additions to the cash balances of the finance companies, and thus a share of current income, used to repay loans, which is not turned obviously to expenditure.

In sum, we have a gross investment demand for consumer finance, I_c, which assumes a position depending primarily on income, and which adjusts itself quickly to an equilibrium reinvestment rate in any income situation. Shifts in this schedule will occur primarily in response to income changes, and as these shifts occur, net investment or disinvestment will take place. But once income stops changing at any level, the equilibrium reinvestment rate is quickly restored. The behavior of this gross demand differs from that for capital goods principally in that it has no intrinsic dynamic tendency except in response to income changes—thus it is not necessarily to be depended on as a regular source of net investment. The rise in popularity of new durable consumer goods, however, which require installment finance, or the shift in consumer habits, may account for independent dynamic shifts in the consumer-finance demand.

Taking this with the gross investment demand for capital goods, we now have two components of the gross investment demand—I_g and I_c. These together constitute the current demand for investable funds to be spent on commodities. Together they may, at a going rate of interest, equal, exceed, or fall short of the combined equilibrium reinvestment rate. With dynamic change stimulating activity in the I_g component, however, they will generally exceed that rate, and there will be some net investment, though varying amounts in successive periods. Both components respond to income changes, so that net disinvestment characterizes periods of declining income and net investment periods of expansion. This combined gross investment demand I $(=I_c+I_g)$, as shown in Figure 55, will constitute, with a minor exception, the entire demand for investable funds to be spent on current output of commodities and services.[20]

[20] Purchase of consumer or capital goods on hand from preceding periods which does not result in current output to replace the stock sets a disinvestment item against the apparent aggregate investment and eliminates such items from

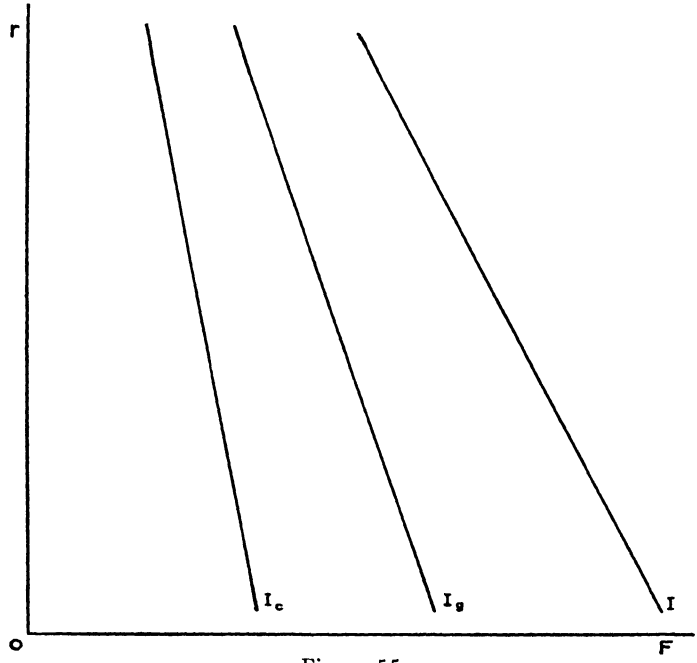

Figure 55

There are, of course, added potential demands for investable funds, resulting from the offering for sale of any existent assets —old securities, capital goods already on hand, and land—by those who own them, which must be added to the combined gross investment demand (I) to obtain the aggregate "demand for investable funds" current at any time. (These yield a rate of interest by selling at a discounted present value of their expected future returns, regardless of their essentially irrelevant original cost, if any.) We will discuss such demands in the following chapter. By and large, however, they are distinguished from investment demand (I) in that they do not lead necessarily to the expenditure of the funds acquired on current output of either consumption or investment goods. Basically, they are simply demands for liquid balances, and although some of the

the true total. Such disinvestment offsets are implicitly deducted in our definition of gross investment demand.

balances so acquired may be spent on current output, preponderantly they will not. Such demands for investable funds thus tend to result in the acquisition of money to be held idle, and they are potentially quite different in significance from investment demands.

With this background of demand for funds, we may now turn to conditions of supply and to the determination of the rate of interest.

SUPPLEMENTARY READINGS

IRVING FISHER, *The Theory of Interest*, New York: The Macmillan Company, 1930, Part II.

EUGEN VON BÖHM-BAWERK, *The Positive Theory of Capital*, London: Macmillan & Company, 1891.

KNUT WICKSELL, *Lectures on Political Economy*, New York: The Macmillan Company, 1934, vol. I, Part II.

JOHN MAYNARD KEYNES, *The General Theory of Employment, Interest, and Money*, Chaps. 11, 12, 16.

ALVIN H. HANSEN, *Fiscal Policy and Business Cycles*, New York: W. W. Norton & Company, 1941, Chaps. XVI, XVII.

WILLIAM FELLNER, *Monetary Policies and Full Employment*, Berkeley, Calif.: University of California Press, 1946, Chap. 3.

CHAPTER 12

INTEREST, MONEY, AND EMPLOYMENT

In discussing the intricacies of investment and the use of capital goods, we have so far dealt mainly with the determinants of the *demand* for capital goods and the corresponding demand for investable funds at various rates of interest. We have not yet explained what the interest rate will be which determines the extent to which any particular demand will be exploited. We have temporarily assumed, however, for purposes of argument, a "given" rate of interest, or in effect, *a perfectly elastic supply of funds* at some arbitrarily fixed interest rate. This would mean that those seeking funds for investment could secure as many or as few funds as they desired without changing this fixed interest charge on money. Although this assumption is not entirely accurate, it is useful for two reasons. First, by neglecting variations in the interest rate, and assuming it to be an arbitrary fixed price, one can demonstrate as we have done that the demands for investable funds for gross investment in capital goods and for consumer loans are quite determinate at any such rate, that, with dynamic change, they may vary systematically over time even though the interest rate is quite inflexible, and that, in stationary conditions, these demands are entirely satiable, so that at any positive rate of interest net investment eventually tends toward zero. All of these things are quite true even though the rate of

interest on funds should approach zero and thus represent a relatively insignificant price.

A second justification for assuming that there is a perfectly elastic supply of funds at some fixed rate of interest is that this assumption does not square too badly with the facts of the modern world. The rate of interest in any strong and wealthy nation today is normally controlled by the national government via the agency of its central bank—in this country by the system of Federal Reserve Banks. The central bank can pretty well set the rate of interest for the economy because it has the effective power to create additional money by making loans or buying securities, or to draw money from circulation by accepting repayment of loans or selling securities. To support a given general rate of interest in the economy, it has in essence merely to announce this rate and then to compete with all other suppliers of funds to make it effective. That is, it would have to make as many loans as were demanded (buy as many securities as offered) at this rate if nonbank funds would otherwise be in short enough supply to lead to a higher rate, or conversely to accept as much repayment of loans as offered (sell as many securities as demanded) at this rate, if nonbank funds would be plentiful enough to cause a lower rate. It can support any given rate of interest by supplying or demanding enough funds to make that rate effective.

Under our system the central bank does make some such rate approximately effective by loan and repayment transactions with commercial banks, by which it regulates the cost of funds to them and the amount of funds they can loan, and by "open-market" transactions in buying and selling securities. It does not deal directly with the bulk of lenders and borrowers in the economy, and thus does not enforce its rate in every locality and on every loan. But it does establish a general "price for money" to, and therefore from, commercial banks, in all principal markets for funds. By its own loan operations it thus sets a general level for the interest rate in the economy, to which the interest on all loans, via a process of competition, tends to adjust. As this happens, there is, in effective approximation, a perfectly elastic supply of funds to the economy at the bank's rate of interest. In wealthy countries with strong governments, the central banking

system is, at least within considerable limits, quite able to support any chosen rate of interest. Furthermore, governments are showing an increasing disposition, for political and fiscal reasons which we need not detail here, to support low and relatively inflexible rates of interest for long periods of time. The student in a hurry could thus do much worse than to enunciate as his theory of the determination of the interest rate that it is set and maintained by the central bank.

If this were a fully correct and adequate statement, our discussion of capital investment might be concluded here—an analysis of investment and the use of capital goods based on the supposition of an arbitrarily given and supported money rate of interest. There are two reasons, however, for going a step further with the matter.

In the first place, the central bank is not the only source of loanable funds in the economy, or the only potential arena for the determination of the rate of interest on money. There are also the cash balances held by individuals and businesses, and the savings of these parties out of income, both of which may be loaned in various amounts at various rates of interest. When the central bank sets and maintains a rate of interest—which it is able to do—other sources of funds must adjust to this rate. Savings and cash balances must reach equilibrium with the bank rate, and the central bank must participate in the process of adjustment as necessary to support its interest rate. There is thus a good deal concealed beneath the plausible observation that the bank sets the rate of interest, and this merits some investigation.

Second, the central bank, by supporting any specific rate of interest, may commit itself either to supplying to or withdrawing from the economy a certain amount of funds, because, at the interest rate it supports, the supply of funds from nonbank sources does not balance the demand therefor. Such additions or withdrawals of money to or from the economy may influence the level of money income and employment. The consequences of the fixing of the rate of interest by the central bank require some analysis. More generally, the whole interrelationship of the supply of funds, the interest rate, and the flow of money income should be considered, and as a part of this consideration the role of the banking system should be determined and evaluated.

To clarify the issues posed it should be convenient to examine the determination of the rate of interest first on the assumption that the central bank and connected banking system is "neutral," or does not participate in the process, and second on the assumption that it does. By neutrality or nonparticipation of the banking system we will mean in effect that the amount of funds or money in the economy is given and fixed, that the banks effectively avoid adding to or subtracting from the amount of money held in the economy. This is the correct definition, since the banks in effect do participate in interest-rate determination only by changing the supply of funds. Our first approximation to a theory of interest will thus suppose a situation where all investable funds are supplied from nonbank sources.

THE SUPPLY OF INVESTABLE FUNDS FROM SAVING

In this circumstance, the rate of interest will in effect be determined by the nonbank demands for and supplies of investable funds, and our main task is to examine the character of these demands and supplies. We have already referred to the principal investment demands and to their determinants, and have seen that they can be precisely expressed only as related to finite intervals of time. The same is true of supplies of funds. A first step, therefore, is to define a unit time interval for which an interest rate is to be determined. This done, we may analyze its determination first for such a period and then over a succession of such periods toward a possible "equilibrium" period.

The unit time period for analysis is conveniently defined as the interval required to complete *one circuit* in circular flow of economic life—for the average "working" dollar in circulation (excluding idle cash balances withheld from circulation) to circulate from income recipients, through expenditure on goods and services, through factor payments, and thus back as income. Each such time period may be defined as beginning just after money income recipients have received a round of money income payments. It thus begins with expenditure out of money income just received, includes the production and delivery of real output corresponding to this expenditure, and ends as the expenditures of the beginning of the period arrive back as money income

payments in the hands of factors of production. The succeeding period begins at this point, and so forth.[1]

The demand for investable funds in any current period depends upon the "going level" of real income, upon whether income has been changing or constant, and upon the level of money prices. By this we must mean that it depends upon the money income and the real income of the immediately preceding period, *as they are projected by expectation* into the current interval, and also upon whether over immediately past periods income has been rising, falling, or remaining at a constant level. The gross demand schedule for investment in capital goods, I_g, then will be made up of a reinvestment component which depends upon the size of the capital stock and its durability and the expected current level of real income, and of a net investment component which depends upon current or recent changes in real income, to which total investment is as yet unadjusted, and upon the rate and character of dynamic changes in techniques and products, supplies of resources, and so forth. The consumer finance demand, I_c, will depend primarily on the expected level of income and the direction of recent movements therein. And both will assume money magnitudes corresponding to the current price level, as inherited from the immediately preceding period. Combining I_g and I_c, we have for any current period a combined schedule of demands for investable funds, I, which shows the amounts of investable funds borrowers will demand at various interest rates in this period for the purposes indicated.

[1] The period is thus defined as the length of time which it takes expenditure, on the average, to elicit output and to become disposable income to persons and to firms retaining gross profits. Reality is artificially simplified by supposing that all income recipients get incomes at a single date rather than in staggered fashion through time, and that all income dollars circulate through the system of enterprises at the speed of the average dollar in circulation. The period may alternatively be defined as the average period of circuit of all money in existence, or that of all money less "idle" money, without altering the principal conclusions of the subsequent analysis; we will here regard the period as the average circuit period of active money, excluding idle cash balances. For purposes of simplified exposition, we will disregard any difference between the circuit periods of consumption and of investment expenditures, tentatively assuming it away.

These are not necessarily the only components of the current demand for funds, but before turning to others, we may examine certain logically parallel sources of supply. These are in effect *savings*, or a supply of funds saved out of income—precisely out of the income received just at the outset of the current period.

Saving for a period of time may be defined for this purpose as the money income available for expenditure in that period less the amount actually spent on consumption (other than consumption financed by consumer loans). The economy receives at the beginning of the period, and has available for expenditure during the period, let us say, $1,000,000. It spends on consumption (before expenditures financed by consumer loans) $800,000. Then it has *saved* during this period $200,000. What is the character of this supply of saving?

For the economy as a whole, gross saving is made up of business saving and individual saving. For any period, business firms as a group will have received a certain gross money income and out of this they pay, or distribute as individual incomes, wages, interest, rents, and a part or all of current profits. The difference which they retain is *business gross savings*, and is made up of accumulations of funds for the replacement of capital goods and of the earnings (interest, rents, profits, etc.), which constitute the undistributed net income of individual owners. Since we include consumer finance companies in the over-all picture, gross saving will include for them repayments of past loans held on hand for reinvestment. This business gross saving for the economy is presumably related to income and relatively insensitive to the rate of interest. Firms will tend to set aside almost automatically as gross saving a reinvestment allotment sufficient to maintain investment (unless net losses prevent this), and ordinarily in addition a portion of the owners' net income for retention in the business. Business gross saving will thus vary with the real income of the economy and also with the money price level as these affect business earnings and reinvestment requirements; on the average it should represent an amount equal to the equilibrium rate of reinvestment plus some net saving of undistributed earnings, which also vary with aggregate income. This saving is probably not very responsive to

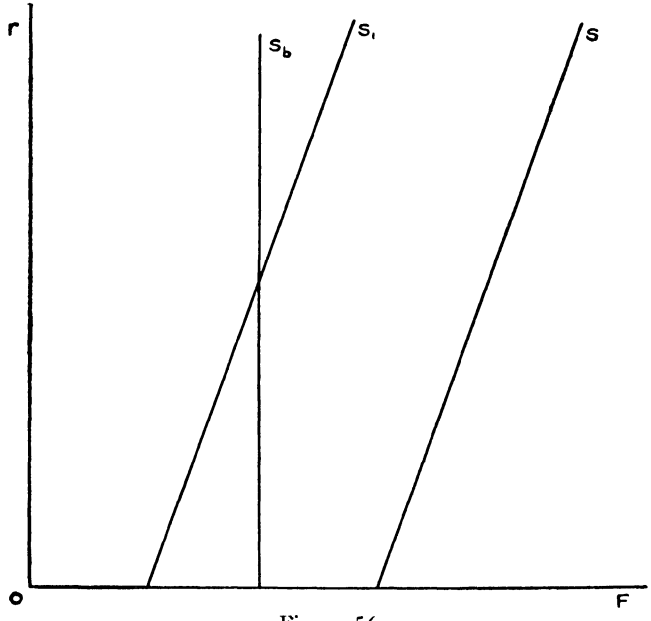

Figure 56

the rate of interest, however; the disposition of business firms to distribute their gross earnings instead of retaining them will not be much affected by interest rate changes.

Coming into any current time interval, then, there is an amount of business saving, representing business gross money income of the preceding period undistributed to individuals. This is an amount depending basically upon the total level of investment and the going level of real income for the economy.[2] It assumes a magnitude as a certain number of dollars of saving when the going money price level is given. Represented as a supply schedule, this saving appears as an inelastic supply unaffected by the rate of interest—S_b in Figure 56. This schedule occupies a given position with the current going real and money income and given money prices; it would shift in positive response to fluctuations in real income, and also would change with changes in the money price level.

[2] And also upon the distribution of income between profits and other shares, as this influences the amount of earnings available for net business saving.

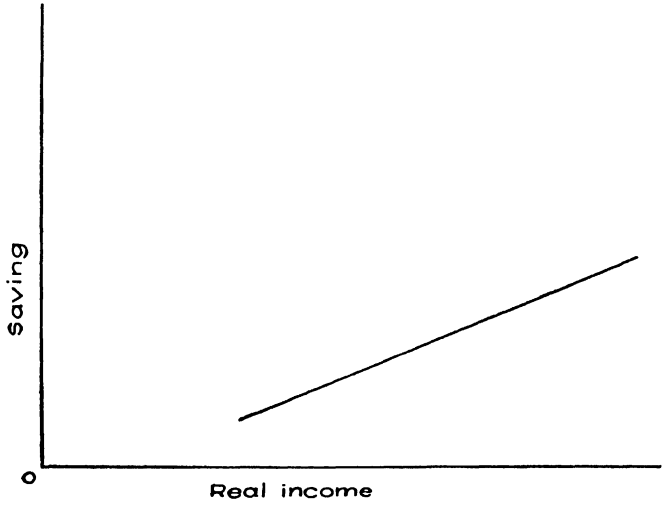

Figure 57

Individual saving is the second component of total saving, and it will represent the share of the money income received by individuals—corresponding to gross income less business saving—which they do not spend on consumption in the period in which it becomes available for expenditure. (Consumption is for this purpose defined as exclusive of that financed by consumer loans.) Such individual saving evidently has a clear positive relation to the real income of the economy, as shown in Figure 57. This relation, however, should be carefully defined. Saving in the sense here employed refers always to an amount of money received at the end of the preceding period, available for expenditure in the current period, and not spent on consumption during the current period. The general relation of such saving to real income is that it will increase with increases in real income and decrease with decreases in real income, provided always that other things influencing saving remain equal—in particular that money prices remain unchanged as real income varies. If money prices are constant, every variation in real income is matched by a corresponding variation in money income, and under these conditions there is a direct relation between money saved and money-and-real-income (variations in money

income exactly reflecting the variations in real income). If real income varies while money prices do *not* remain constant, this means essentially that a definite change has occurred in the value of the economy's output *as measured in dollars of constant purchasing power* (*i.e.*, adjusting for changes in the money price level), whereas no necessarily corresponding change in money income has occurred. Then it holds that there is an identical direct relation of money saved as measured in dollars of constant purchasing power—*i.e.*, of the purchasing power of money saved—to value of income measured in dollars of constant purchasing power (to real income, or to purchasing power of money income). The basic relation indicated in Figure 57 is thus one of money saving as measured in "adjusted" dollars, or of real purchasing power of money saving, to real income, or to real purchasing power of money income. It becomes a simple relation of unadjusted dollars of saving to real income only on the special assumption of a constant money price level.[3]

At any given interest rate, individual saving in dollars of constant purchasing power will tend to vary with variations in real income. At any given real income level for the economy, such saving will have a determinate "adjusted dollar" magnitude and, with the going money price level, will represent a corresponding number of actual dollars. With any current level of real and money income, there will be a corresponding amount of individual money saving. This amount may be somewhat responsive to the rate of interest, in that people will be induced to save more by higher interest rates, even though their saving depends primarily on income. Thus we may represent a supply schedule of individual saving for the current period, out of last period's income, as a schedule S_i in Figure 56. This has some positive elasticity to interest-rate changes, and will shift (along lines indicated by Figure 57) with shifts in income.

Business gross saving plus individual net saving make up total saving for the current period. Schedules S_b and S_i are added to get a combined saving schedule, S (Fig. 56). This schedule shows the amounts of the going money income, as received from

[3] The preceding is emphasized at length because saving represents money or money purchasing power not spent on consumption, and *not* in any direct sense an amount of real goods saved (which may differ from money saving).

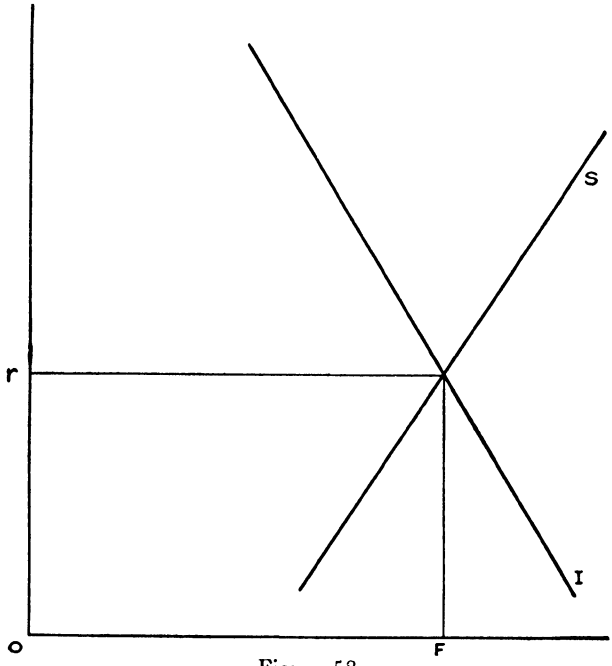

Figure 58

the previous period, which will be "saved"—*i.e.*, not spent on consumption [4]—during the current period.

This is a supply of funds received as income which, since it is not spent on consumption, becomes potentially available for expenditure on investment—that is, to fulfill the demand for investable funds. It is also a supply which is potentially somewhat responsive to changes in the interest rate. Do we now have enough information to determine the interest rate for a single period? We have a combined demand schedule for investable funds, I $(= I_g + I_c)$, and a corresponding combined supply schedule of saving, S $(= S_b + S_i)$. Can we determine the interest rate at the intersection of these two schedules, as in Figure 58, at the rate of interest r? It would be simple if this were so, but in general it is not. The rate of interest is not necessarily so determined as to equate gross investment and gross saving in any

[4] Consumption excluding expenditure financed by consumer loans.

current period. The reason for this is that savings are not automatically or necessarily invested; the fact that income from last period is not spent on consumption goods does not mean it must go into investment. In effect there may be for the economy as a whole *net hoarding* of currently saved money, or *net dishoarding* of previously saved money (idle cash balances), so that investment may exceed or fall short of saving. To explain this possibility we must refer to certain additional demands for and supplies of funds which we have not yet taken into account.

CASH BALANCES, HOARDING, AND LIQUIDITY PREFERENCE

The crux of the explanation lies in the existence of *cash balances,* in the fact that they can be increased or decreased, and in the probability that in certain frequently recurring circumstances people will wish to increase or decrease them. At any time there is some amount of money in circulation, and by money we mean currency, coin, and bank credit, or whatever is freely accepted in payment for goods. (We have supposed, by holding banks hypothetically neutral, that this amount of money is provisionally fixed for the economy.) This money circulates as payments for commodities and factor services through the economy, resulting in a flow of money income. But not all of it circulates all the time. Individuals and business firms throughout the economy regularly hold balances of money in "idle" form—maintain certain average unspent balances of money. Thus John Jones has a monthly salary of $500, which he regularly spends for various purposes. But he maintains a minimum cash balance of $750 more—his cash account never falls below $750, and fluctuates between $1250 immediately after he receives his salary down to $750 just before he receives it again. If this is true for John Jones, it can be true for the economy. But what does it mean for an economy with only a fixed amount of money (M) to hold? A given state of idleness of balances for an economy essentially means either that a certain share of total M is not in active circulation or alternatively that there is a drag on the speed of or income velocity of circulation of the total money supply—on the average rapidity with which it circulates from income payments through the productive system and back to in-

come payments again. When we say, therefore, that an economy holds a certain amount of idle cash balances, we mean that a part of the available money is not circulating regularly or, interchangeably, that the total money supply is circulating at a certain restricted income velocity and thus supporting some certain corresponding amount of money income per period of time.

The economy then holds "idle balances" in a certain quantity at any given time. It may obviously increase or decrease its idle balances if it so desires. It may add more of the money supply to idle balances—*i.e.*, hoard—or it may withdraw previously idle money from balances and spend it. Or, to put it in alternative form, the economy may spend a given total money supply more rapidly or more slowly than it has been doing. Suppose an economy has 60 billion dollars of money in existence. In a given current state, 20 billion of this is held idle or noncirculating, and 40 billion circulates. The circulating or working money has, we will suppose, a circuit velocity of three times per year—that is, it passes from gross income of enterprises around through the economy and back to gross income of enterprises three times annually. Gross money income per year is thus 120 billion (40 times 3). In this circumstance the average income velocity of all money is 2, or 120 divided by 60. Now the economy may add to its idle balances—suppose it adds 10 billion. Idle balances are now 30 and circulating money 30. If the circuit velocity of the latter remains at 3, money income will be 90 per year. Or, alternatively, the average income velocity of all money is 1½ or 90 divided by 60. It is thus established that the economy with a given total money supply can hoard or dishoard cash balances, and also that hoarding or dishoarding will tend to decrease or increase money expenditure and income.

The third significant observation is that individuals and businesses holding cash balances may wish to increase or decrease them. Under certain circumstances, that is, people may prefer an addition to their cash holdings to more securities or other noncash assets, and thus may hoard. Under other circumstances they may prefer to give up some of their previous idle balances in return for more securities or other assets and thus may dishoard. What is the rule governing this sort of behavior? The ruling motive has been aptly described as the psychological

attitude of *liquidity preference.*[5] Money is *the* liquid asset, freely exchangeable for all goods, and no other asset or security "can make this claim"—to use other assets for spending power they must first be converted into cash. This liquidity attribute of money gives it a certain premium over other assets for individuals and business firms—other things equal, they would rather hold money than other assets. They have, moreover, certain definite demands or needs for liquidity—for working balances, as "reserves" against emergency expenditures or unforeseen reverses, and also, speculatively, against the possibility of a fall in the price of securities or assets.

But the need for liquidity is not a need for a certain absolute amount of money; it is a variable need or desire, which can be more or less completely fulfilled, like the desire for more suits of clothes or more theater admissions. Balance holders have not only a liquidity preference for cash balances, but a marginal rate of liquidity preference for additions to cash balances at any given point. This marginal rate of liquidity preference may be represented as the interest rate return which balance holders are willing to forego to secure a small addition to any given amount of balances, or which they demand if they are to give up a small amount of balances in return for claims on future payments in the form of securities or other assets. In any given income situation, the marginal rate of liquidity preference decreases as balances are increased, showing additions to balances are worth successively less in interest return foregone. The rational balance holder will then adjust his cash holdings so that his marginal rate of liquidity preference is equal to the going rate of interest return on assets—so that the psychological "yield" of liquidity is on the margin equal to the cost of having it in terms of interest return foregone.

If this is so, there will be a negatively sloped demand schedule for liquidity, showing, for any given income situation, the amounts of balances which people will hold at each rate of interest. There will be additions to cash balances as the interest rate falls below a certain rate and decreases in cash balances as it rises above this level. It appears in effect that an economy of

[5] For the basic development of this notion, see J. M. Keynes, *op. cit.*, Chaps. 15 and 17.

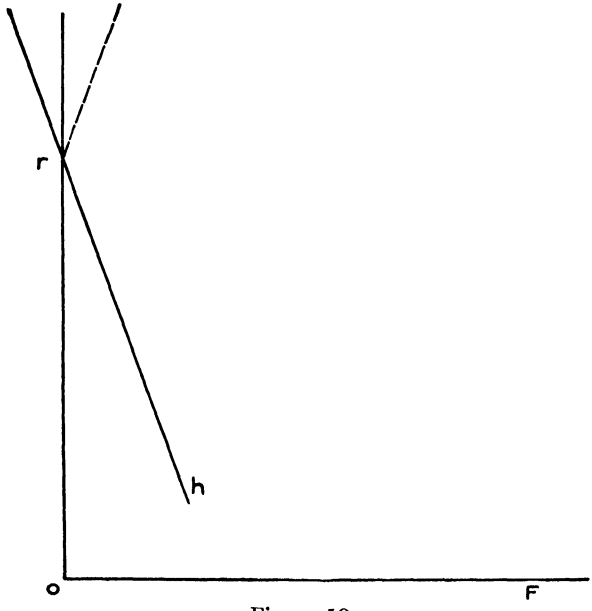

Figure 59

such people, having a desire to hold cash balances which is related to the rate of interest, may hoard (add to) balances if the rate of interest is "too low," will hoard less if it rises, will hoard nothing at some intermediate rate, and will dishoard (subtract from) balances above this rate. *This is because they increase or decrease their idle cash holdings to bring their marginal rates of liquidity preference into balance with any going rate of interest.*

The net effect of hoarding on the supply of investable funds—in addition to or in subtraction from current saving in any period—can be summarized in a single "hoarding" (or liquidity) demand curve, as in Figure 59. The net hoarding curve *h* shows that in a given time period cash balances would be held constant at the rate of interest *r*, that below this rate there would be net demand for hoarding in the amounts shown, and that above this rate net dishoarding (or an addition to the *supply* of funds, as shown by the dotted line) would take place in the indicated amounts.

Such a schedule describes over-all behavior without indicating its components, however, so we may clarify the matter a bit if we break this net hoarding curve into its major parts. Let us look in turn to hoarding by savers, by holders of cash balances, and by holders of securities and other assets. Savers, by definition, are those who have received an income from the previous period and have not spent all of it on consumption. This saving adds to their cash balances at once, but it is not automatically invested. In effect the savers can either invest their saving, exchanging it for interest-bearing capital or other assets or for securities giving claims thereon, or they can hold it in the form of cash. Following the liquidity preference logic, there is some interest rate at or above which they will invest all their savings. But below this rate they will have a demand schedule for hoarding, showing amounts of their saving they will add to their balances at various rates of interest. Now let us turn to balance holders—persons or firms which hold certain cash balances from the preceding period. Below a certain rate of interest, they will surrender none of these balances—will invest none of them in interest-bearing assets or claims. But above this rate they will supply certain increments of funds from these balances, thus giving rise to a positively sloping supply curve of funds from balances. Since savers and balance holders, however, are in general the same group of individuals and firms, let us aggregate the preceding two tendencies. Now we have it that persons and firms holding balances and saving may either add to their cash balances, or hold these balances constant, or decrease their balances—by investing in or offering for assets less than, the same as, or more than, saving. What they do will depend upon the rate of interest, and with variations in r we get from savers and balance holders a supply schedule of investable funds, DH (gross dishoarding) like that shown in Figure 60.

This schedule reads as follows: At some rate of interest r, savers and balance holders as a group will in the current period invest just what they save, holding their balances constant. Below this rate, they will hoard (as shown by a negative supply), subtracting so much from the supply of saving. Above this rate they will make available the additional amounts of investable funds shown, as they adjust their cash balances to the rate

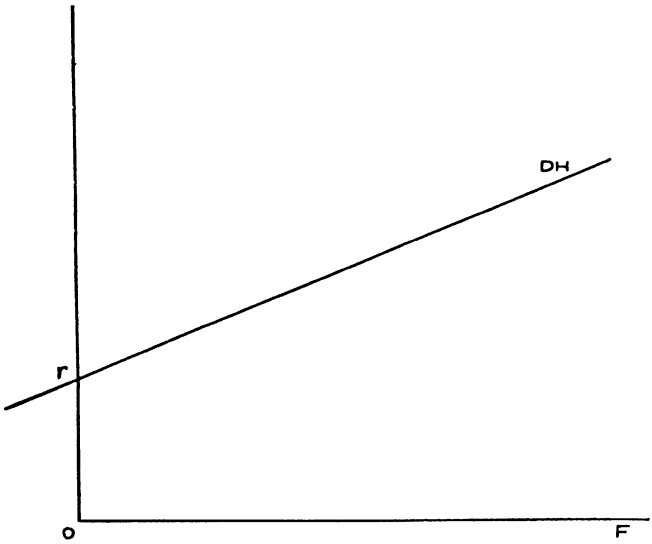

Figure 60

of interest. We thus get an additional supply curve of investable funds, *DH*, a supply of old balances or of dishoarding, positively related to the rate of interest. This supply, however, becomes negative below a certain interest rate.

The further component of the liquidity picture is the demand for cash balances by those holding securities—those "old" securities which have been issued in conjunction with past investment and represent interest-bearing claims on those investments. These securities are the principal "old assets" which people may wish to be able to exchange for cash, but may be considered as supplemented by any other salable assets, including land and old capital goods inherited from previous periods. Holders of these assets will be desirous, according to liquidity-preference calculations, of exchanging them for cash in increasing amounts below a certain rate of interest,[6] and thus will create a demand

[6] How may the "rate of interest" on securities already in existence, and potentially bearing some "fixed" rate of return be said to vary? The answer is simple—by variations in the selling price of the securities. Thus suppose that the supply of old securities is made up entirely of bonds of $100 principal value, repayable in 20 years and bearing a return of 3 percent of $100, or $3

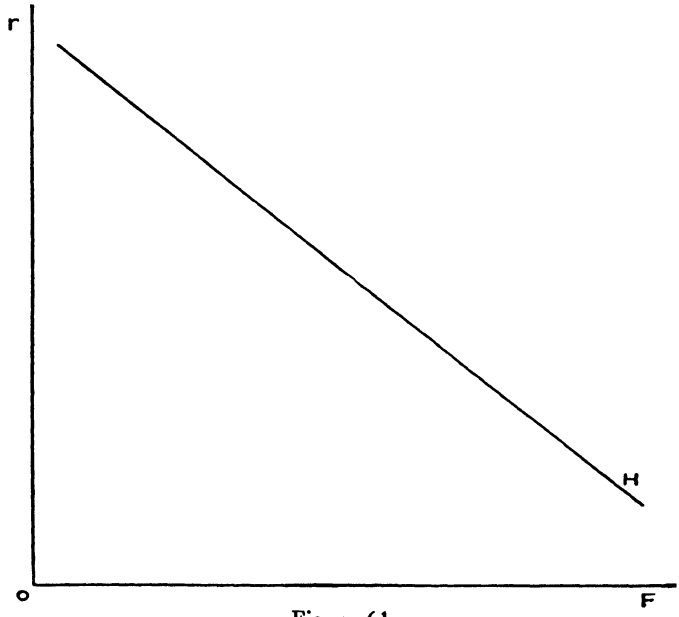

Figure 61

for funds of the sort shown in Figure 61. We label this demand curve as H because it represents in effect *gross hoarding—i.e.*, the demand for cash balances to be held in idleness in lieu of earning assets. (It is possible that some of the funds acquired by selling old securities may be turned to current consumption—such a fraction of these funds as are so shifted from idleness we will neglect, or in effect consider transferred from the H, or hoarding curve, to the consumer-finance fraction (I_c) of the I or investment curve.)

Taking the gross dishoarding curve DH, which represents the dishoarding of balance holders and savers, and the gross hoarding curve H, together, we have the component demands for balances for hoarding and supplies of balances by dishoard-

per year. For these securities to yield 3 percent, they must sell at $100. But they may be sold at a higher price to yield less than 3 percent, or a lower price to yield more than 3 percent. Variation in the rate of interest on old securities of fixed yield is thus accomplished by a variation in their prices, and fixing of an equilibrium market rate on them is a matter of fixing their prices.

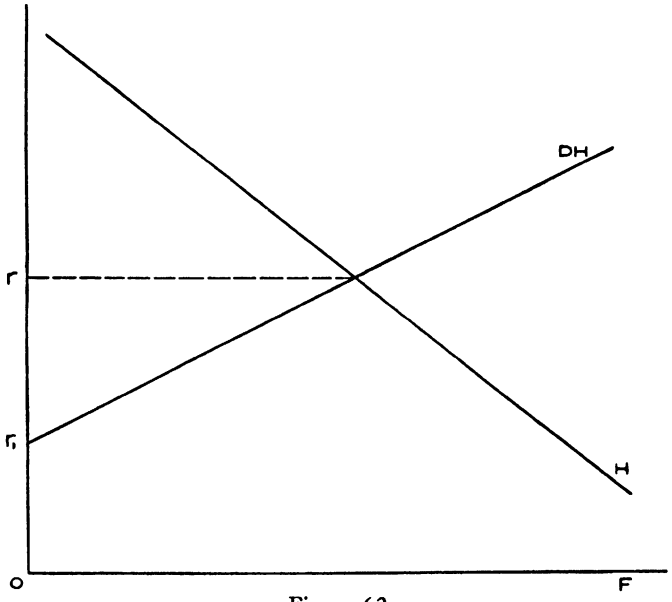

Figure 62

ing, which together determine the net hoarding propensity of the economy in any particular period. If we consider them together, as in Figure 62, it is apparent that if the interest rate were at the level r, where the two schedules intersect, there would be no net hoarding or dishoarding of balances in the economy—as much would be hoarded by sellers of old securities as was dishoarded by holders of balances. If the rate of interest were above r, there would be net dishoarding, since the sellers of old securities would take less to hoard than was dishoarded by holders of balances. Or in other words, the *total* supply of funds available for current *investment* would exceed saving above this interest rate. If the rate of interest were less than r, there would be net hoarding in the economy, since the sellers of old securities would demand more to hoard than was dishoarded by balance holders. (Below the rate r_1, there would be additional net hoarding by savers.) Therefore, below the interest rate r the supply of funds available for investment is necessarily less than saving, because of net hoarding. It will also be ap-

parent that if the supply amounts in the *DH* schedule were subtracted from the demand amounts on *H* at each rate of interest, we would get a net hoarding schedule like that shown in Figure 59. What we have done is thus to break the net hoarding schedule into two component parts in order to take explicit account of the exchanges of old securities which are a part of the process under examination.

It will be noted that these schedules express marginal rates of liquidity preference, for hoarders and dishoarders, in any current situation—their preferences for perfectly liquid cash over securities and other assets. This liquidity preference is not based on the *risk* attached to securities—risk, that is, of default on principal amount or interest payments. Such a risk may lead to a risk premium which is added to the yield of securities, but that is a separate matter.[7] We refer here to the preference for liquidity as compared to riskless securities, as based upon the desire to hold cash for transactions, for contingencies, and in speculation against a future rise in the rate of interest. The positions of these liquidity-preference schedules, moreover, will depend not only upon people's habits and attitudes relative to the holding of cash balances but also on the relative abundance of cash to satisfy their liquidity needs. The relative abundance of cash, in turn, will depend upon the ratio of the amount of money in circulation and the level of money income per period of time. Idle cash balances will be larger or smaller as the money supply is greater or less in relation to money income. The positions of the *H* and *DH* schedules will therefore depend upon the supply of money, the size of money income, the volume of securities in existence, and people's liquidity attitudes.

THE RATE OF INTEREST FOR A SINGLE PERIOD

How, now, does the introduction of these tendencies to hoard or dishoard, as represented in these demand and supply schedules for cash balances, influence the determination of the interest rate? We may consider this question first for a given time period, where there is a given amount of money in circulation and

[7] This is discussed on pp. 425-429 below.

where a certain money and real income has been inherited from the preceding period. In this period, the interest rate will evidently be determined at such a level as to equate the aggregate demand for and the aggregate supply of investable funds. But the aggregate demand for investable funds includes not only investment demand I, but also a demand for liquid balances H; and the aggregate supply of investable funds includes not only savings, S, but a supply of liquid balances, DH. And, what is most important, the interest rate which equates the aggregate demand and supply need not be such as to produce equality between saving and investment. There may be net hoarding or dishoarding. In fact, there will be one or the other if the interest rate which would make for no hoarding (where H equals DH) is different from the interest rate which would equate saving and investment (where S equals I). In this event, the equilibrium rate as determined by aggregate supply and aggregate demand will fall between these levels, and there will be a discrepancy between investment and saving, positive or negative, which will be balanced by an identical amount of dishoarding or hoarding of balances. [As we proceed with the discussion from this point we will for brevity drop certain modifying expressions previously attached to the term "investment"; and by "investment" (unmodified) we will refer to I as previously defined, or, in other words, to "gross investment expenditure of the current period," including current net investment plus current reinvestment both in capital goods and in consumer finance. This is the relevant investment concept in analyzing the demand for and use of funds in any current period.]

This is easily demonstrated as follows. In any time period, we have a given investment demand schedule, I, representing amounts of funds to be invested and spent for output in this period in addition to consumption, and also a given saving schedule, S, representing amounts of money income not spent on consumption and thus available for investment or hoarding in this period. *Alone*, these schedules would determine an interest rate r_1, as in Figure 63, and saving and investment would be equal. Further we have a demand schedule for funds, H, representing amounts of funds to be demanded in this period not for spending but for additions to balances, and a supply schedule

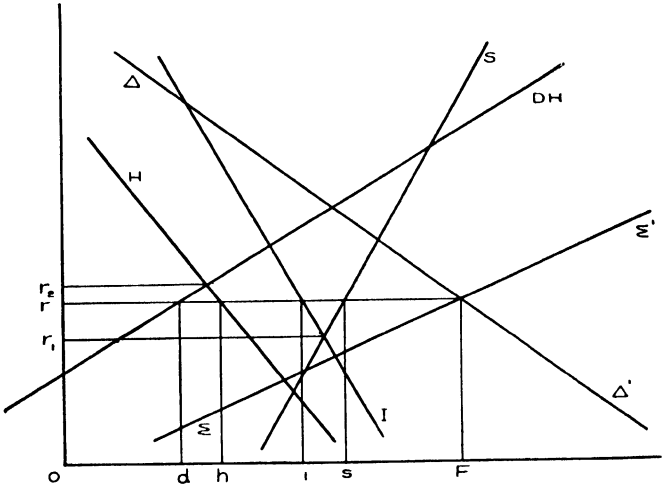

Figure 63

of funds, DH, showing amounts of funds to be supplied from cash balances in addition to saving. This latter pair of schedules *alone* would determine an interest rate r_2, as in Figure 63, an equilibrium marginal rate of liquidity preference between balances and old securities. The aggregate demand for funds, however, is investment plus hoarding demand $(I + H)$, which may be represented as $\Delta\Delta'$. And the aggregate supply of funds is saving plus dishoarding $(S + DH)$, or $\Sigma\Sigma'$. The equilibrium interest rate for the period is evidently determined at the intersection of the Δ and Σ schedules (Fig. 63) and this rate, r, necessarily falls between r_1 and r_2. At this rate of interest, although the aggregate demand and supply for funds will balance, it is not necessary that investment, I, be equal to saving, S, nor that gross hoarding, H, equal gross dishoarding, DH, since the interest rates which would balance the respective demand-supply pairs are not necessarily identical. In the case illustrated, for example, savings (os) is greater than investment (oi) by the amount is, at the equilibrium rate, and this is exactly counterbalanced by the difference, dh, between gross hoarding (oh) and gross dishoarding (od). In effect, an amount equal to the excess of saving over investment is hoarded in the net, by savers

and by sellers of old securities, as their demands for liquidity are sufficient to keep the rate of interest above a level which would equate S and I.

In general, the discrepancy between S and I may be in either direction, or they may, coincidentally, be equal. A somewhat larger investment demand (I) in the preceding example would bring investment equal to saving, at a higher interest rate, and result in no net hoarding. A still larger investment demand would make investment greater than saving and result in some net dishoarding of cash balances. The strategic reason for possible differences between saving and investment lies in the existence of idle cash balances, which may be drawn upon or augmented, and in the liquidity preferences of those holding balances, which will induce them under certain circumstances to dishoard and under others to hoard. The rate of interest at any given level of income thus does not operate to equate saving and investment.

Although the reasons for our repeated emphasis on the possibility of the inequality of investment and saving is presumably obvious, it may be useful to restate this reason explicitly. If the rate of interest were set at such a level that *saving* (*i.e.*, amounts of money income not spent on consumption goods) would be exactly equal to *investment* (*i.e.*, amounts spent on output in addition to consumption), then the total expenditure on output of each new period (which equals consumption plus investment) would be exactly equal to the money income of the previous period. Total money income would always be exactly spent for output, on one thing or another, since all savings would be so spent.[8] Since expenditure on output generates income, this in turn would mean that money income would remain *constant* from period to period through time. This would be true because there would be no additions to or subtractions from cash balances. If investment is different from saving, however, *then money income will change over time*—that is, the expenditure and resultant income of the period in which the discrepancy

[8] Any amount of consumption expenditure which actually buys goods produced in previous periods and not currently replaced results in a disinvestment item deducted in computing total investment; hence the sum of consumption and investment is equal to expenditure on current output.

occurs will change from the income of the preceding period. If investment exceeds saving, money income will increase; if investment is less than saving, money income will fall. (The increase or decrease is matched by a fall or rise in cash balances.) Money income will in effect move from period to period *until* it reaches a level at which saving equals investment. And concurrently, prices, real income, and employment will also tend to adjust, since changes in money income must induce corresponding changes in either prices or physical output.

This is important for two reasons. First, out of the interaction of saving and investment there may be determined, after a number of periods, an equilibrium level of money income, and possibly of employment. The process of money income, price, and employment adjustments must be traced to determine the character of such an equilibrium. Second, the rate of interest itself cannot be viewed as finally determined until money income has reached a stable resting place. So long as saving and investment are unequal from period to period, continued income movements will produce changes in the rate of interest, and the equilibrium value of this rate cannot be said to be fully determined until income reaches an equilibrium value.

THE BANK RATE OF INTEREST

Before turning to these issues, we should add further to our analysis of interest-rate determination in any given time period. To this point we have argued on the assumption that the amount of money (which we may designate as M) was fixed in the economy. By so doing we have abstracted from the operation of the banking system, the essential function of which is to supply money, in the form of bank credits, to the economy. It is recognized that the banking system, as governed or influenced by the central bank, may either create new money, which is added to the circulation, or retract the supply of money. Money or credit creation is accomplished by the banks by making loans or by buying securities, in return for which they set up credits or "deposits" which pass as money; retraction of the money supply is effected by "calling" loans for repayment or by selling securities, thus canceling out an equivalent amount of credits or de-

posits. The banks are thus a potential source of investable funds, and also, alternatively, act as potential agents in taking such funds off the market. How do the banks affect the determination of the rate of interest?

The general effect of bank action is revealed by inquiring what the effect is upon the interest rate in a given period if the banking system succeeds in increasing or diminishing the supply of money. Suppose the current equilibrium interest rate, with given M, is at the rate r shown in Figure 63 above. An increased amount of money would obviously reduce this rate. With more money in circulation the liquidity needs of persons, relative to the going money income, would be more fully satisfied, and therefore the supply of funds from balances DH would shift outward and down and the demand for hoarding H would be decreased. This would result in a fall in the rate of interest and a new equilibrium with larger investment and smaller saving. Correspondingly, a decreased M would mean that liquidity needs were less fully satisfied; the supply of balances would be decreased and the demand for hoarding increased, and the rate of interest would tend to rise. By changing the amount of money in circulation, the banking system may thus affect the rate of interest and also the relation of saving to investment.

The banking system does not operate, however, by arbitrarily adding or withdrawing money from the economy. It operates by charging a bank rate of interest and by supporting this rate and thus making it effective for the whole economy. It does this by loaning or accepting repayment of loans, or buying or selling securities, in whatever amounts are necessary to support its rate. More specifically, the central bank—which is represented in this country by the Federal Reserve System—establishes and supports such an interest rate, making it roughly effective for the various commercial banks of the system by regulating the amount of "reserves" that they hold and the interest rate at which these reserves can be acquired, as well as by participating directly or indirectly to some extent in the security markets. It is thus generally true, though subject to numerous exceptions of detail, to say that "the bank"—*i.e.*, the governmentally controlled central bank—sets the rate of interest for the economy.

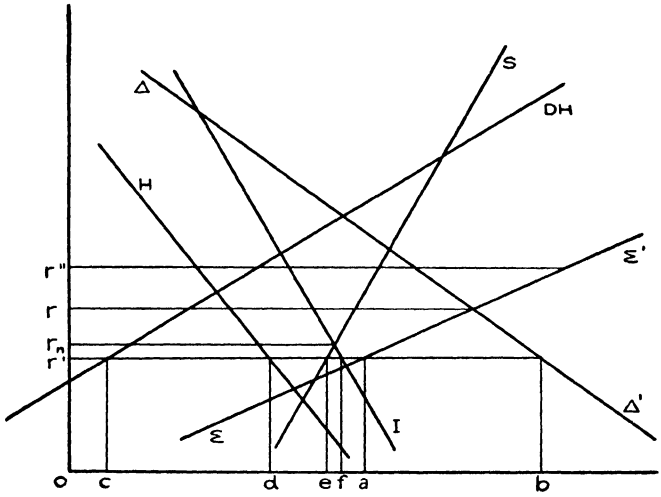

Figure 64

Let us investigate this process to see what it implies. In an economy with a given income and provisionally given money supply, we will suppose for one time period the existence of the primary determinants of the current interest rate, *I*, *H*, *S*, and *DH*, as shown in Figure 63 above. We reproduce such a family of demands and supplies of investable funds in Figure 64. The interest rate which will prevail in the absence of any bank borrowing or lending is here *r*, where the aggregate demand and supply for funds is in balance. Now if the banking system sets *r* as its own rate, there will *currently* (at this income) be no net movement of funds in or out of the banks—they will be called upon neither to buy nor sell securities, neither to loan nor recall loans, in the net, and will in effect pursue a "neutral" policy. Suppose, however, that by design or accident the banks are not neutral, and set a lower or a higher rate. If the banks set and support a lower rate *r'*, they will be called upon, in the current period to which the demand and supply schedules refer, to supply funds by loaning or by buying securities, in the amount *ab*, the discrepancy between the supply and demand for investable funds from nonbank sources at this rate. In effect, they would have to add the amount *cd* to the net hoarding of cash

balances, plus the amount *ef* to investment, in order to sustain the rate r'.

If I and S are unequal at r', as shown, income will, of course, move, and the bank will face a somewhat altered situation in the next period. If the bank sets a higher rate than r, at r'', the opposite situation prevails. The supply of funds from nonbank sources will exceed demand, and the bank will be obliged to buy securities or retract loans in the amount of the difference, thus restricting the amount of money in circulation. In general, the banks can support any rate of interest as a fixed price in the economy so long as they are able and willing, from period to period, to buy or sell sufficient securities to support this rate. When the rate of interest is fixed by the banks, the amount of money becomes a variable which adjusts to the prevailing conditions of demand and supply for funds.

It will be well to remember, however, that we have referred explicitly so far only to a single time period of bank operations. A succession of periods with changing income and a changing amount of money introduce additional problems. While we are considering a single period, however, we may as well inquire whether the bank may not, by setting its interest rate appropriately, *stabilize* income by equating saving and investment. In the case illustrated in Figure 64 above, for example, cannot the bank set its rate at r_n, at which S would be equal to I, and thus insure a stability of income from this period to the next, from that period to the following, and so forth? Assuming that the bank wishes to do so (in this case to prevent a decline in income), it should be able to stabilize income from one period to another so long as it can supply loans or buy securities sufficiently to meet all unsatisfied liquidity demands at this rate of interest. Conversely, if I tends to exceed S, at the equilibrium rate of interest (r) as determined by exclusively nonbank demand and supply, and a higher rate of interest is required to equalize them, the bank can support such a rate and stabilize income so long as it is willing and able to sell securities demanded at this rate, sufficiently to relieve people of the "excess" liquidity they hold. Now is it reasonable to suppose that the bank is generally *able* thus to stabilize income, against increases or declines, if it so chooses?

The answer is not necessarily affirmative. As for reducing the interest rate to correct a situation where S exceeds I, the bank may be unable to do this, either for the given period or over a succession of periods. If I and S can be equated only at a negative or zero rate of interest—if they intersect only at or below a zero rate—the bank will be practically unable to equate them, since zero or negative interest rates are evidently inconsistent with a controlled supply of money. Similarly, it may also be unable to equate S and I if the interest rate at which they balance is quite low, and if at this rate the demand for liquid balances becomes insatiably large. It is possible, that is, that below some certain positive interest rate the demand for liquid balances (by those who would borrow from or sell securities to the bank) becomes *very elastic*, so that the bank would be unwilling or unable (perhaps because of legal reserve requirements) to supply enough funds to support this rate. Regardless of how much money it pumped into the system from period to period, the demand for balances would still exceed the nonbank supply, and the bank would eventually reach the limit of its resources and have to raise the rate. Thus unless the saving and investment schedules are sufficiently interest-elastic to intersect at some rate above such a positive "psychological minimum" rate, the bank may be unable to make them balance and to prevent the decline of income. There is thus a definite theoretical limit on the power of the banking system to curtail the downward movement of the level of income.

Suppose, conversely, that in the current balance as determined without bank participation, I exceeds S. Is the bank always able to raise the interest rate sufficiently to bring savings and investment into balance? It may again be powerless if at very high rates of interest there is a very large supply of liquid balances which it will be called upon to assume, by selling securities or by calling loans. Thus, if the supply of balances becomes *very elastic* above a certain interest rate, the bank may not be in possession of sufficient callable loans and salable securities to support a higher rate. Then if I and S intersect above this rate, the bank will be literally unable to forestall a rise in income, unless the government is willing to create high-yield securities specifically for the purpose of obtaining funds to impound or to

withdraw from circulation. There is thus in effect also some maximum interest rate which the banking system can ordinarily support, and this also potentially limits its power over the movement of income. It should be noted, however, that the upper *theoretical* limit is less likely to be tested in practice than the lower, if only because investment is likely to respond in a rather elastic manner to very high rates of interest.

Within the limits indicated, the banking system, as regulated by the central bank, is in a position to set the rate of interest, and when it does so, *the interest rate becomes a fixed price, and the quantity of money becomes a variable.* How can one characterize the rates of interest ordinarily set by a banking system, like our own Federal Reserve System? The determination of the general level of interest rates is a matter of deliberate governmental policy, ordinarily designed to create favorable business conditions—*i.e.*, with reasonably full employment, stability, and an avoidance of price inflation. It follows that, within limits, bank rates will tend to be adapted to the current level of income and employment. In periods of low employment, the aim of governmental policy will ordinarily be to generate an increase in money income, which may be expected, if money factor prices are fairly rigid, to induce a rise in employment and output. The appropriate bank rate in this situation is a very low one, to stimulate investment and restrict saving as much as possible. Cyclical depressions and periods of stagnant employment will thus ordinarily find a low rate of interest supported by the central bank. In periods of rising employment, the bank rate may be adjusted upward somewhat to limit the excess of I over S and to retard the rate of rise of income. When full employment is reached and if income is still rising (investment exceeding saving) a high rate of interest is generally indicated. In conditions of relatively stable income and high employment, the bank will ordinarily attempt to manipulate the rate of interest to maintain an equality between S and I, and to decrease the interest rate when income begins to fall off. The rate of interest under modern central bank conditions is thus likely to respond, because of policy decisions, to the current business situation.

There are, of course, definite limits on the efficacy of bank interest policy to accomplish the desired results on income. In

the first place, as we have noted, there are theoretical maximum and minimum limits upon the bank rate of interest, set by the character of liquidity preferences. These limits may be narrowed if there are legal limitations on the amount of credit which the banking system is allowed to create. Further, it must be remembered that the central bank can only *guess* at any current time what the proper rate of interest is—it may well err in this regard in spite of the best intentions.

Three other practical limitations should also be noted. First, the banking system can fully control rates on all types of securities only so far as it deals in all of them. If it is limited to buying and selling only short-term securities, its influence on the yields of long-term securities will be only indirect. Then loans for long-term purposes may carry different rates than the bank rate if the market anticipates rising or falling interest rates in the future. Second, the central bank, as an agent of the government, may find it inexpedient to change the interest rate—especially to raise it—because of the effect of such increases on the government debt. A marked rise in the rate of interest would depress the price of publicly held fixed-yield government bonds, and might thus lessen confidence in the government. Finally, since increases in the interest rate adversely affect those holding all fixed-yield securities, so far as they may wish to sell them (a rising interest rate means falling security prices), the unpopularity of an increase in interest rates above conventional levels may operate to discourage such an increase. In practice, therefore, and especially when a large government debt is held by the public, the bank rate of interest may tend to be rather inflexible at some level not exceeding the yield rate on government securities.

We thus return to our initial observation that the rate of interest in general is determined by the central bank. To this we may add that although in theory it may be set at various rates depending on the level of employment, it tends increasingly in recent times to be set at some relatively low level and held at or near that level rather inflexibly over time. In this circumstance, the interest rate becomes a relatively fixed price (at some level high enough to forestall indefinitely large demands for liquidity) and the demand and supply for funds in any period is

balanced by the bank's adding to or subtracting from the money supply sufficiently to preserve the balance at its interest rate. The supply of money, as provided by the banks, thus becomes a variable which is adapted to varying conditions of demands for funds for investment and for liquid balances, and of supplies of savings.

THE CONDITIONS FOR MOVEMENT AND STABILITY OF MONEY INCOME

Let us return to the problem of money income and its movements. Before examining bank policy, we saw that in any initial period of time, with a certain going level of money and real income and with a given amount of money, a balance of the demand and supply of investable funds would be struck. In this balance, saving would not necessarily equal investment. For any one time period—defined as average time interval required for money to circulate once around the circular flow of the economy—the amount of money saved and thus subtracted from the flow of income need not be equal to the amount of money invested and thus added to the flow of income. Business plus individual gross savings do not necessarily equal the total of gross expenditure on capital goods plus borrowing to finance consumer expenditure. This would be true even though the money supply were fixed and the interest rate freely adjustable; inequality of S and I is possible by virtue of net hoarding or dishoarding of cash balances.

The supposition of a fixed money supply and a freely moving interest rate may now be replaced with the more realistic assumption that there is a relatively fixed bank rate of interest and a correspondingly flexible money supply. In this situation also the amount of money saved in any initial period need not be equal to the amount invested. This would be true, for example, in Figure 64 above, if the bank set any rate other than r. In general there will be a given investment demand I and a given saving supply S, which intersect at some rate of interest r_n, as in Figure 65. This rate of interest may conceivably be high, low, or negative. The demand for and supply of idle balances may now be provisionally neglected. In this situa-

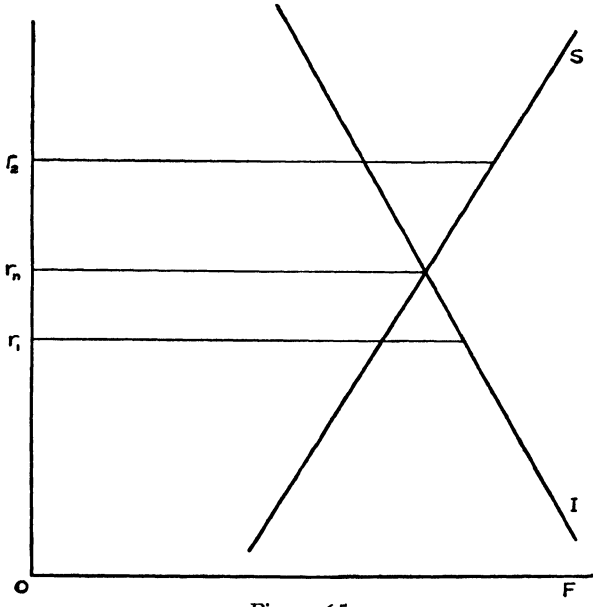

Figure 65

tion, the bank rate may be viewed alternatively as being set equal to or greater or less than the rate at which saving will equal investment. A higher or lower rate might be set by design, or by accident, or because of the inability of the bank (theoretically or practically) to set the rate r_n. In any event, in the absence of the ability plus the inclination of the banking system to support a rate equal to r_n, S will not be equal to I in any initial income situation. Let us view the rate of interest as set by the banks alternatively at r_n, r_1, and r_2, in Figure 65, and consider the consequences.

Support of the bank rate at the level r_n will result in an equality of S and I and therefore in the stability of money income from the first period to next. The reason for this is that gross saving subtractions and gross investment additions to income are just in balance at r_n. Suppose that the total gross income flow for a period as it arrives as sales receipts of enterprises is 80 billion dollars. Out of this, business gross savings are 10 billion, so that enterprises pay as income to individuals

70 billion. Out of their 70 billion personal incomes, individuals save 5 billion at the rate r_n and spend 65 billion on consumption. Total savings are thus 15 billion. If now exactly 15 billion of funds are devoted to investment spending—let us say reinvestment of 8 billion, net investment in capital goods of 4 billion, and net consumer finance additions of 3 billion—an additional 15 billion of expenditure will be generated. Total expenditure of the period will be 65 billion consumption out of income plus 15 billion investment, and *therefore* the sales receipts of enterprises in the succeeding period will again be 80 billion. The money income flow will have remained constant over the two periods because saving is balanced by investment.

Suppose, however, that the bank rate is set at a lower rate, r_1, perhaps because the bank wishes to generate a rise in income, or perhaps because it is practically unable to raise its rate. Then I will exceed S. Business plus personal saving is lower, let us say, at a total of 10 billion, while investment of all sorts is higher, at 20 billion. Now the consumption expenditure from the 80 billion dollar gross income flow is 70 billion, and is augmented by an investment expenditure of 20 billion. The income of the succeeding period will rise to 90 billion. An excess of I over S causes a rise of income, accompanied almost certainly by an increase in the amount of money in circulation as the banks meet the difference between the total demand for funds and the total supply from nonbank sources. Conversely, a bank rate set at r_2 would result in an excess of S over I and a decline in money income. Such a rate might be set deliberately, or because I and S intersected at a lower rate than the bank could maintain. If business plus individual savings were higher at 20 billion, and investment lower at 10 billion, consumption out of 80 billion initial income would be 60, plus 10 of investment expenditure, so that money income in the succeeding period would decline to 70 billion.

Generalizing, it is obvious that if, at the going interest rate, S equals I, money income remains constant; if S is greater than I, money income falls; if I is greater than S, money income rises. Inequality of S and I causes money income to move; equality allows it to remain stable. Money income will move *until* saving becomes equal to investment. Moreover, a free-

moving rate of interest will not prevent such money income movements, and a bank rate of interest may by necessity or by choice be set so as to allow them.

THE EQUILIBRIUM OF MONEY INCOME

Our first step has been to observe the conditions which will begin a movement of money income. Where does this movement end? If the money income from the end of period 1 is 80 billion, and if in period 2, 70 billion is consumed and 20 billion invested, the income of period 2 becomes 90 billion. But what of periods 3 and 4—what is the saving-investment relation there? Can S and I be expected to become equal after a succession of income movements, and if so at what point?

The analysis from this point forward is complicated by the fact that movements in money income (set off by inequality of S and I) may cause changes in both S and I, and (in the absence of fixed bank policy) in the interest rate. Supposing the interest rate to be held fixed by the bank throughout, we must take account of the response of the investment demand schedule and of the saving supply schedule to the movement of money income. What is the character of these responses? Movements in money income will cause I and S to change because of the resultant changes in employment and real income in the economy and the resultant changes in money prices.[9] For short-term analysis account must also be taken of investment changes based on the anticipation of or speculation on further real income or price changes, but we will neglect this here. When money income increases (because I initially exceeds S) this must induce a rise either in employment or in money prices or in both, and a decline in money income will elicit a decline in either or both. It will be convenient to isolate the effect upon saving and investment of an induced price change and of an induced change in employment and real income. Let us first abstract from the

[9] The resultant change in money prices involves a change in money factor prices, and some corresponding change in money commodity prices, which may also involve some change in the relation of commodity to factor prices and in profits.

effect of money income changes on prices, and center attention entirely on their effects on real income and employment.

In effect, let us assume an economy with a given *rigid* money price level, which will not change in response to changes in money income. This would involve given money factor prices and given price-cost relationships for commodities. (We will provisionally neglect any changes in price-cost relationships which are implicit in income movements.) The assumption is logically tenable so long as money income does not continue to rise after full employment is reached. In this economy we begin with some initial money income, to which there corresponds, at given prices, a certain level of employment and real income, inherited from the immediately preceding period. At this money and real income level, the current period has a given supply schedule of money savings and a given money demand schedule for investment, and also a bank rate of interest, at which saving and investment are by supposition unequal. Money income therefore changes, money prices remaining fixed and employment changing in response. How, now, will money saving and investment respond to the resulting identical changes in money and real income? This is a useful question, first because in many situations money prices are in fact relatively rigid and also because answering it may allow us to determine the *net* effect of *real* income movement on saving and investment of funds.

The response of saving to a change in real income is thought to be fairly predictable; that is, money saved as measured in dollars of constant purchasing power will be positively associated with real income (which is money income as measured in dollars of constant purchasing power).[10] Saving so measured will increase as real income increases and decrease with real income decreases. Gross saving so measured tends at any given rate of interest to follow a relationship to gross real income about like that described by the line S_g in Figure 66 (where S is gross saving in dollars of constant purchasing power and Y is gross real income). That is, at some very small level of income, *ob*, the economy saves nothing, being so poor that it must consume its entire current gross income. Business saving for reinvestment

[10] This relation was discussed on pp. 372-374.

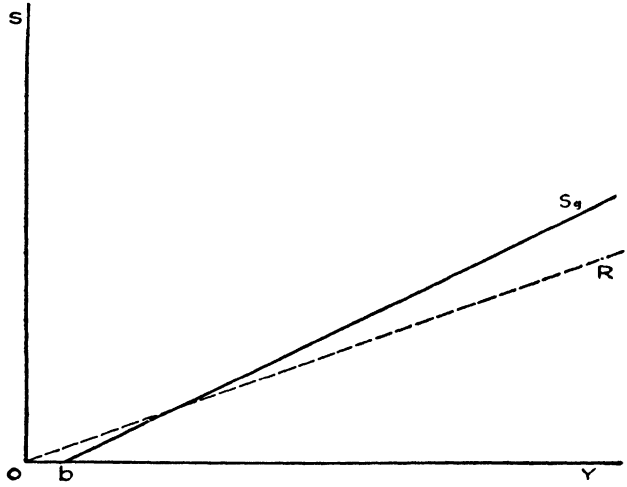

Figure 66

is thus insufficient to maintain the stock of capital goods at the level reached through past investments, and individual net saving is zero or less; their sum, gross saving, drops to zero. Above this minimal level, each increment in real income will result in a corresponding but smaller increment in gross saving, roughly in a straight-line relationship, so that the ratio of an increment in saving to a corresponding increment in real income remains constant for successive changes in income.[11] This increase in saving reflects both an increase in business saving for reinvestment and an increase in net saving from net income. As this net saving increases, gross saving becomes successively larger than that required for reinvestment alone. Thus the dotted line R in Figure 66 might represent the rate of business saving necessary for an equilibrium rate of reinvestment (just to maintain the capital stock) at various levels of income, on the supposition that the total stock to be maintained increases directly with income. To the left of the intersection of S_g and R, gross saving is less than the equilibrium rate of reinvestment, net saving is negative, and the capital stock is consumed to some extent. To the right of this intersection, net saving in the economy

[11] It is possible that saving increases at an increasing rate against income, so that the line S_g would curve upward.

really begins, as gross saving then surpasses the equilibrium rate of reinvestment; from this point outward, an increasing quantity of net saving seeks outlet in net investment. The response of gross saving to income is thus compounded of a response in saving for reinvestment plus a response in net saving.

So much for the presumed relation between saving as measured in dollars of constant purchasing power and real income. Under the special simplifying assumption of constant prices adopted in this section, the purchasing power of the dollar cannot vary. As a consequence actual dollar amounts of saving do not have to be adjusted for a changing value of the dollar, as every change in real income is implemented by an identical change of money income. In this special context we may say simply that there is a direct relation of money saving to real income of precisely the sort shown in Figure 66 and described in the preceding paragraph. The relation of unadjusted money saving to money-real income is identical to that which holds more generally between adjusted saving and real income.

Under these assumptions, the response of the supply of gross money saving to changes in the level of money and real income is illustrated in Figure 67, where the gross saving supply schedules for three successively larger levels of income are represented by S_1, S_2, and S_3, and the corresponding equilibrium reinvestment rates by R_1, R_2, and R_3. The excess of total saving over normal or equilibrium reinvestment becomes larger as real income increases.[12]

The response of investment to the same income change may be analyzed along similar lines. The gross investment demand, as a demand for a number of dollars to be spent on current capital-goods output, represents a certain basic real demand for a given physical quantity of goods, which is reflected in some specific money demand determined by the going level of money prices. Gross investment demand will respond to changes in real income in a presumably positive fashion, in the sense that the amount of real output demanded for investment will increase as real income increases and decrease as real income decreases,

[12] Cf. A. H. Hansen, *Fiscal Policy and Business Cycles*, Chap. 11, for a discussion of the saving-income relation.

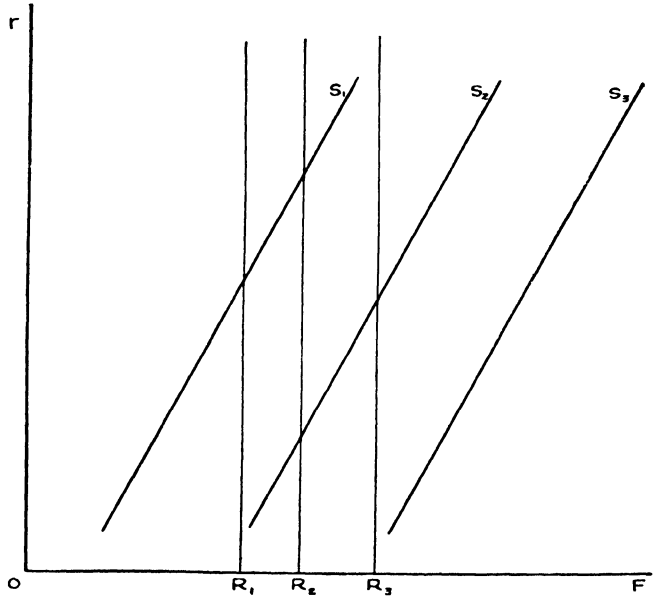

Figure 67

which also means that the demand for investable funds as measured in dollars of constant purchasing power will respond positively to changes in real income. (If the money price level changes, the actual money investment demand will vary in response to this change, but the investment demand in adjusted dollars will not necessarily be affected.) Under the special assumption that the money price level is constant, however, no adjustment need be made for price changes, and it follows that there is a direct relation between actual dollars of money demanded for investment and the level of real-and-money income.

Following this assumption, let us look at reinvestment and net investment demand for funds. The reinvestment demand in any initial situation will depend upon the level of past investment in capital goods (and in consumer finance) and will tend to equal the rate of reinvestment saving. As money-and-real income increases, and with it the accumulation of past investments, reinvestment demand tends to rise proportionally, keeping pace with reinvestment saving. As far as reinvestment goes,

therefore, the demand for and supply of funds tend to remain roughly in balance despite increasing income.[13]

The rate of net investment, or position of the net investment demand schedule, depends in any initial situation upon the accumulated opportunities offered by technological change, product development, and so forth. Now net investment demand will increase in response to increasing money-and-real income for two reasons. First, any going opportunities for net investment at the initial level of income will be greater at a higher income level. This will account for some shift in the investment demand schedule in response to increasing income. Second, all investment will tend to be increased as money-and-real income increases, to adapt the stock of capital goods (and consumer finance) to a higher level of output. This will cause an added net investment demand *while income is increasing*, but this source of net investment demand will vanish once income stops increasing and simply remains at a higher level. Thus, at a money-and-real income level of 100 billion, the economy may require 200 billion as a total accumulation of investments in capital goods, and at 150 billion income it may require 300 billion of accumulated investment. While the economy moves from 100 to 150 billion income, then, there will be a net investment demand of 100 billion. Once it is stabilized at the new higher income level, there will be no such net investment demand remaining.[14]

An increase in money-and-real income will thus cause any existing net investment demand to be somewhat larger, and at the same time create an additional but quickly satiable net investment demand to adjust the supply of capital goods to the larger income. As a whole net investment demand will respond positively to changes in income, but part of this response will be transitory and tend to vanish when income approaches stability at a higher level. Net investment is thus positively related to

[13] We provisionally neglect short-term discrepancies of reinvestment demand and reinvestment saving, which should be taken into account in more detailed treatments of the theory of employment.

[14] See Hansen, *op. cit.*, Chap. 12, for a discussion of the relation of investment demand to change in income.

income, but, except for periods of income *movement*, it is relatively inflexible.

Considering reinvestment and net investment together, the net investment gap between gross investment and reinvestment is not likely to widen so much, as money-and-real income increases, as does the gap between gross saving and reinvestment saving. Net investment does not respond to income changes, except transitionally, as much as does net saving—hence gross investment is less responsive to income than gross saving. As money-and-real income increases, net saving inevitably increases and remains up, but net investment increases less, or at any rate only temporarily, so that it can *maintain* a smaller increase than net saving.

This brief and somewhat sketchy survey of the relation of investment to income suggests that if we were to trace the movement of income in detail, we should have to take into account short-run *fluctuations* in investment engendered by income movements as well as longer run ultimate shifts in investment. It is not our intention here to examine this and other aspects of the theory of fluctuations. We will therefore be content with the *general* observations on the relation of saving and investment to movements in real income:

1. that gross saving will increase and decrease directly with increases and decreases in money-and-real income.
2. that gross investment will also increase and decrease with money-and-real income, but ordinarily by smaller amounts.

Let us now return to our initial question—the process of adjustment of money and real income (with given money prices) to an initial inequality of saving and investment. Suppose that, in a given situation and with the bank rate of interest r, gross investment (I_1) exceeds gross saving (S_1), as in Figure 68. Money and real income therefore increase. But as income rises, period by period, saving increases, and at the same time investment increases, though on the average at a lesser rate. Thus in a second period the schedules shift to I_2 and S_2, where the investment still exceeds saving, but by a smaller amount. Income continues to rise until, in some period 3, saving (S_3) becomes equal to investment (I_3). When this point is reached, money-

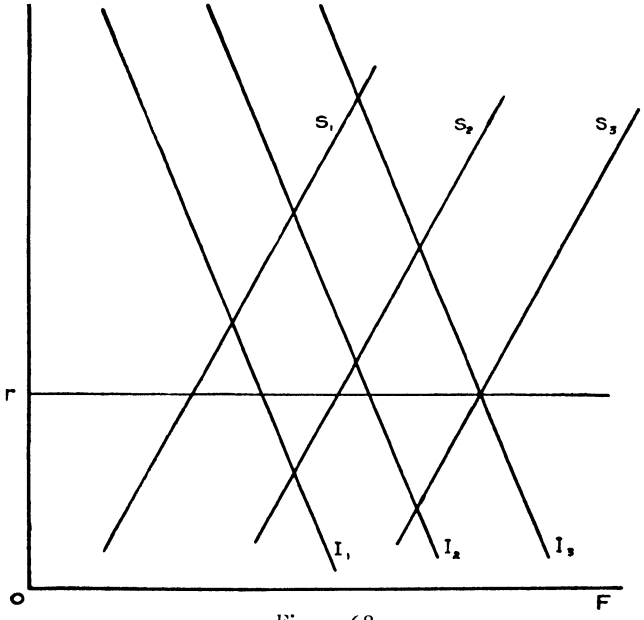

Figure 68

and-real income will tend to be stable so long as investment demand remains at the same level.

In effect, money-and-real income moves until saving out of income becomes equal to investment. This is equally true of situations where there is an initial excess of saving over investment—here income will decline, and with it saving (rapidly) and investment (less rapidly), until at some lower income saving is equal to investment.

We thus have two essential propositions. One, which is obvious, is that an inequality of saving and investment will result in a movement in money and real income *until* saving and investment become equal. The second, perhaps less obvious, is that some movement in money-and-real income (given money prices) will tend to bring them into equality, because saving is more responsive to real income than is investment. From the two propositions we derive the idea of *an equilibrium level of real income,* which real income would tend to reach with any given set of money prices. Given such prices, money and real

income tend to move to this equilibrium level, where saving out of income is equal to investment.

The preceding may be illustrated by means of an arithmetical example in the form of a "model sequence" development of income through time. Suppose money prices are absolutely given, and also the rate of interest. Suppose, also, that the rate of income (Y) is initially \$100 per period, and that the rate of investment (I) is \$50 per period. Suppose that at the \$100 level of income, people consume \$60 and save \$40. This gives us initially an inequality of saving and investment. As income moves, let us suppose: (1) that investment is totally insensitive, remaining at \$50 per period, but (2) that saving changes with real income so that saving is always 40 percent of the income of a period. Now, letting Y stand for the gross income received as sales receipts by enterprises at the beginning of each period of time, I for investment, C for consumption, and S for saving, we can trace the movement of income, period by period, toward equilibrium.

Period	Y	$C(= .6Y)$	$S(= .4Y)$	$I(= 50)$
1	100	60	40	50
2	110	66	44	50
3	116	69.6	46.4	50
4	119.6	71.8	47.8	50
5	121.8	73.1	48.7	50
6	123.1	73.9	49.2	50
n	125	75	50	50
$n + 1$	125	75	50	50

Beginning with the first period, with an income of 100 available for expenditure, consumption is 60 and investment 50, creating an income of 110 available for expenditure in period 2. Out of the latter income, consumption (after 40 percent saving) is 66, and with investment remaining at 50, total expenditure, and hence the income for expenditure in period 3, is 116. We then progress from period to period with income rising as long as

investment exceeds saving. But with stable investment (entirely insensitive to income), saving overtakes investment as income rises, until in period *n* income approaches the level 125. This is the income out of which saving equals investment—in this case, the income 40 percent of which is equal to $50. When this income is reached, no further income movements will occur so long as (1) investment remains at 50 per period, and (2) people's saving habits remain unchanged. Thus the income 125 is an *equilibrium level of income* (given certain money prices)—income moves to this level, which is defined as the income the saving from which is equal to investment. There is always potentially some such equilibrium level of income corresponding to (1) the rate of investment, and (2) the relation of saving to income.[15]

If, of course, investment is not totally insensitive to income but responds in some degree to income movements, the process of adjustment will be prolonged and the equilibrium level of income altered. In the sequence just developed, for example, the eventual equilibrium might involve an investment per period of 60 (up 10 because of higher income) and an income of 150. But the general principle holds that, with given money prices, income will move to that level at which saving equals investment.

Now that this general principle has been stated, let us examine a few of its implications. First, it is true that, on the assumption of given money prices, the equilibrium level of money income involves some corresponding equilibrium of real income and hence an equilibrium level of employment. What we have said in effect is that in this case *employment and output move to such a level that saving equals investment.* By abstracting from price movements, we have made all money income movements equivalent to movements in real income and employment, and

[15] The student will note, for example, that in the preceding sequence, if investment were 40 per period, the equilibrium income would be 100; if it were 30, the equilibrium would be 75; if it were 60, the equilibrium income would be 150. Conversely, if investment were 50 and people saved 25 percent instead of 40 percent of their incomes in general, the equilibrium income would be 200 instead of 125.

have argued toward an equilibrium level of employment. The basic determinants of this equilibrium of employment are the relation of saving to real income and the behavior of real investment demand. Assuming given money prices, we explicitly trace the adjustment to equilibrium of money income in response to money saving and investment. But the saving and investment adjustments rest basically on real income changes (as opposed to price changes), and the equilibrium money income reached carries with it a corresponding equilibrium level of employment.

Second, this equilibrium level of employment may be less than full employment. Full-employment income can be sustained only if the saving from such income is not greater than the currently corresponding rate of investment. Suppose that at a level of income sufficient to sustain full employment (*e.g.*, a money income of 200 with given prices) the economy will save 25 percent (50 in money at the same prices). Then investment must be at least 25 percent of income (50 per period) at the full-employment income of 200, or that level of employment cannot be sustained. If investment will not rise above 20, employment and income will decline to that level where saving is 20, and involuntary unemployment will occur. Why cannot full employment *automatically* be sustained? There are two reciprocating reasons. First, investment does not automatically emerge in sufficient amounts to absorb the savings from income. The general rate of investment depends in considerable part on the net investment demand for new capital goods, and this in turn emerges mainly in response to dynamic change in techniques, products, and so forth. In the absence of a sufficient pace of such change, net investment demand may be small and gross investment thus relatively reduced in amount, for any level of real income. Second, the rate of saving is linked in definite fashion to the level of real income. With the given bank interest rate assumed, the economy will save a relatively high proportion of a full-employment income, and diminishing amounts from smaller incomes. Thus a shortage of investment opportunity does not lead to a restriction of saving *at a high income level;* it leads rather to a reduction of income. With a reduced investment demand, the economy cannot sustain full employment be-

cause it insists upon saving "too much" of its income at that level of employment.[16]

A third implication of the general principle is that an excess of investment over saving at the income where full employment is reached means that the assumption of rigid money prices is no longer tenable. In effect, if such a discrepancy continues at full employment, a rise in money income must continue and will be matched by a corresponding rise in money prices—by inflation. It should be noted, however, that a balance of investment and saving may very often be reached short of full employment, so that the *necessity* of changing money prices is not encountered.

Fourth, an investment demand which changes over time will tend to induce a changing level of employment. New net investment demands arise principally from dynamic changes in techniques and the like, but each additional demand is satiable and as time passes must be supplanted by others. With a varying pace of dynamic change in the economy, investment demand and the rate of investment will fluctuate. If there is a relatively fixed relation of saving to income, then income and employment will tend to fluctuate also. In periods of a higher rate of investment, employment will tend to be high; if investment is small, employment will be low. Given the relation of saving to income (and assuming given money prices), the rate of investment determines the rate of employment.

Fifth, movements of money income as it adapts to a changing state of investment demand (given money prices and with a given bank rate of interest) will be accompanied by variations in the amount of money in circulation, as the banks buy or sell securities to maintain the balance between the total demand for and supply of investable funds. Thus a rising money income and employment will be matched by some corresponding rise in bank loans or by security purchases by banks, and a decline in income

[16] The argument as presented explicitly assumes a fixed rate of interest. It should be noted that it may apply equally well with a variable rate of interest. That is, interest rate reductions may not balance saving and investment at full employment income because (1) saving and investment may respond in rather inelastic fashion to interest rate changes (*i.e.*, net hoarding will continue in spite of low rates), and (2) the total demand for liquidity will tend to prevent the reduction of the interest rate beyond a certain positive minimum.

by some decline in bank loans or by security sales by banks. It is not necessarily true that the exact difference between investment and saving be supplied or retracted by banks in each period, since there may be also some net hoarding or dishoarding of balances. But this difference will be in large part counterbalanced by expansion or contraction of the money supply.

It is worth noting in this connection that even with a fixed supply of money an equilibrium level of employment would be struck in much the same fashion, so as to equate saving and investment. Income would still move until these two became equal. But with a fixed money supply, the interest rate would also move as the ratio of money income to the money supply varied (thus changing marginal liquidity preference)—it would rise with an increasing income and fall with a declining income. This adjustment of the interest rate would cooperate with the adjustment in income to equalize saving and investment, and would thus tend to reduce the amplitude of income movements. Money income changes would then be accomplished entirely by variations in the velocity of circulation of the given money supply.

Summarizing our argument to this point, we have pointed out that, with given money prices, saving need not necessarily equal investment at all levels of money or real income. This is true either with a fixed supply of money and a free interest rate, or with a fixed interest rate and a variable money supply. If saving and investment are unequal, money income, and with it real income and employment, will tend to move until they become equal. There is then some *equilibrium* real income and employment, the level of which is determined by the relation of saving to real income, and by the size of investment demand and its response to real income changes. With given money prices, money income and employment move to such a level that saving equals investment. This level may (or may not) involve involuntary unemployment. With the bank rate of interest also given, the equilibrium level of money income will be matched by some corresponding amount of money as supplied by the banks.

This covers the adjustment of money and real income on the supposition that there are given and fixed money prices. On this

assumption, it is possible to reach fairly definite conclusions concerning the simultaneous equilibrium of saving, investment, and income. In what respects is our analysis altered if we recognize that money prices may rise under the influence of increasing money income and fall with declining money income or because of severe unemployment? This is an important question, since, until the effect of money price changes is admitted, a *true* equilibrium of money income and of employment is hardly established.

MONEY PRICE CHANGES AND THE LEVEL OF EMPLOYMENT

The principal impact of money price changes upon the determination of money income and employment stems from their effects on money investment and money saving. A general rise or fall in money prices will cause some change in the amount of money investment corresponding to a given real investment demand, and in the amount of money saving corresponding to a given real income. If we are initially in a period where income is 100, consumption 80, saving 20, and investment 30, money income necessarily rises, reaching a level of 110 in the following period. If this rise in income results entirely in a price increase (of 10 percent), real income and employment and the interest rate remaining the same, and if the real demand for investment goods remains the same, money investment should now be 33. If people's saving habits relative to real incomes are unaltered, saving should be 22. The rise of money income would then continue, since money investment and money saving have responded directly to changes in the price level.

We may expect in general that a rising money price level (rising in response, ordinarily, to an increase in money income) will elicit some corresponding upward shifts in money investment and saving, and that a declining money price level will have the opposite effect. As the economy moves through time, and as money income is changed because of inequalities of saving and investment, the movements of money income and employment are much complicated by this fact. Money prices are ordinarily not rigid, and money saving and investment thus change in response *both* to real income changes *and* to changes

in the price level. When we can no longer assume rigid money prices, we must appraise in detail the effects of changes in these prices.

Precise appraisal of the effects of money price changes on investment and income requires that we distinguish between the effects of a changed price level per se, and the effects of anticipations of further price change, as they may arise from changes already experienced. A higher general price level, if it is expected to be maintained, tends to cause an upward adjustment in the money dimension of everything—more money is paid for given goods, a higher money wage corresponds to a given real wage, and investment and saving are similarly affected. Except so far as the *structure* of prices is altered by alteration of the price level and if the interest rate is inflexibly given, the changed price level may thus imply nothing further than a changed money income. There is no necessary *resultant* alteration in output or employment, and no necessary resultant change in the *relationship* of saving to investment, even though the money dimension of both may be increased. If this is true, a higher price level per se (with given interest rate) does not alter any pre-existing tendency of employment to rise, and a lower price level does not overcome a pre-existing tendency of employment to fall. Changing price levels accentuate the tendency of *money income* to change but tend to be neutral in their effects on employment, *so long as we overlook anticipation of further price change, or speculation on price changes.*[17]

This is the position taken by J. M. Keynes [18] when he holds in effect that, speculative matters aside, the money price level is neutral in the determination of equilibrium employment—that real income and employment will move to a determinate level regardless of the level of money prices with any given interest rate. This is essentially because employment depends upon the relation of saving to investment, and because money price changes tend to affect both saving and investment proportion-

[17] This is provided the supply of money is perfectly elastic at the going rate of interest, so that a changing price level does not alter the interest rate or the marginal rate of liquidity preference—a situation which holds as long as the bank supports a certain inflexible rate of interest.

[18] *General Theory of Employment, Interest, and Money,* Chaps. 19-21.

ately, leaving their relationship unaltered. If in a given situation investment is invariant at 40 per period, income is 100, and saving is 20 percent of any income, then in the absence of any money factor price changes, money income would rise to 200 per period. At this income, saving would become 40 and equal investment. This 200 of money income we will suppose sufficient to support an employment of resources of 1000; thus the equilibrium employment is 1000 with given money factor prices. Suppose, instead, that as money income rises from the initial situation just described, money factor prices also rise for some reason, eventually doubling their previous values. This will tend to cause money investment per period to rise to 80, since the same real investment will require twice the money at doubled capital goods prices. Money income will therefore rise to 400, at which money saving will equal 80. But at doubled money factor prices, an income of 400 will support only 1000 of employment, and the employment equilibrium is unaltered. The only effect of the rising price level is to accentuate the increase in money income, and, at a given bank interest rate, to elicit the creation of more credit money. The determination of the equilibrium level of employment and real income is *as if* there had been a rigid money price level—only *money* income levels and movements are affected by price changes. The economy tends to move to a determinate level of employment regardless of what happens to the money price level.

The application of this argument to falling money prices is perhaps the most interesting. We have already indicated that with given money prices an underemployment equilibrium can occur. That is, the rate of investment can be insufficient to absorb saving from full-employment income, and income and employment decline until saving does equal investment. The question naturally arises whether a fall in money prices can now increase the level of employment. Consistent with the preceding argument, the answer is no. A decrease in prices causes a proportionate decline in money investment; hence money income declines proportionately to the drop in prices, and employment is unaffected. So long as there is some given level of real investment demand (so that money investment simply varies directly with price changes) and so long as there is some given relation

of saving to real income, money price reductions will not cause an increase in employment. They will lead only to declines in money income, and, with a fixed bank rate of interest, to a contraction of the money supply.

This argument is perfectly valid as far as it goes. Higher or lower money price levels per se will not affect the level of employment, unless they influence the real investment demand or the fundamental relation of saving to real income. In saying this, we emphasize that money price changes *do not* influence employment levels. Perhaps it may be well to emphasize also that they *do* influence money income changes. Thus the rise in money income generated by an initial excess of investment over saving would ordinarily be fairly limited if money prices were rigid. But if money prices rise with rising income, the increase of money income is accentuated and may proceed to a much higher (theoretically unlimited) point. Correspondingly, where investment is less than saving, successive declines in money prices may lead to an indefinitely declining money income. The consequences of a rising money price level are particularly significant when such a rise is virtually forced by an excess of investment over saving after full employment has already been reached. Of central importance, however, is the tendency of employment and real income to move to an equilibrium level largely regardless of associated movements in money prices.

This general conclusion concerning the effect of money price changes is potentially deficient, however, in three respects. It overlooks the effects of speculation on price change; it assumes a given bank rate of interest; and it neglects associated changes in the price structure.

Speculation on future price change may, of course, have a considerable effect on the current amount of investment. The expectation of higher money prices over some future interval will generally stimulate current investment, as buyers acquire capital goods before their prices rise. This applies both to durable equipment, like machinery, which may be bought "in advance," and to inventories, which may be bought and held for a price increase. Conversely, the expectation of lower money prices over some future interval will reduce current investment, as the purchase of capital goods is postponed. It follows that a part of the

immediate importance of money price changes, as they result from changing money income, will stem from the anticipations they create with respect to further changes.

These effects, however, are not always easily predicted. When prices are rising with rising money income, some anticipation of further price increases is ordinarily created, and this will tend currently to induce a greater than proportionate increase in money investment; thus a 10-percent price increase might lead to a 20- or 30-percent increase in money investment because of advance buying. Conversely, a decline in prices may lead to a greater than proportionate decrease in money investment. So far as this is true, money price movements, as they respond to money income movements, will tend to accentuate the current discrepancies between investment and saving and to lead to more severe and prolonged movements of both money income and employment. On the other hand, a continued rise of money prices to an unaccustomed high level may create the anticipation of an eventual decline, and a decline of money prices to abnormal lows may cause investors to anticipate an eventual price rise. When these "contrary" expectations become operative, price movements tend to narrow the gap between investment and saving and to arrest further movements of money income and employment. It is partly because of such contrary expectations that a progressive inflation or deflation of money income and prices may tend ultimately to be self-stopping. It is also for this reason that severe declines in money factor prices may provide some stimulus to investment and employment. Our conclusions concerning the neutrality of the price level in the determination of equilibrium employment must thus be modified to allow for the effects of "speculation" on investment.

A second assumption limiting these conclusions is that of a fixed bank rate of interest and its concomitant, a flexible money supply. This is not a serious limitation, since the interest rate is maintained rather inflexibly by the banking system under most circumstances. If, however, the money supply should be rather inflexible beyond certain limits, so that the interest rate moved freely, the following amendments to our argument would have to be introduced. Rising money prices, generating additions to money income, would put increasing pressure on money balances

and cause the rate of interest to increase, thus putting some drag on the increase of investment and income. Falling money prices would release money for balances and depress the rate of interest, thus tempering the decline of investment and income. The amplitude of movements in money income and employment is thus less augmented by price changes when the money supply is fixed than when it is flexible.

Could falling prices increase employment, via the increase of liquid balances, except by depressing the interest rate? Not unless the accumulation of balances alters the relation of saving to real income, causing people to save a smaller proportion of a given real income. There is some disagreement on this point, but we will suggest here that saving will not be much altered by liquidity, except via changes in the interest rate. In sum, the assumption of a fixed interest rate and variable money supply is not far from the facts. So far as the facts are otherwise, our main conclusions are not seriously altered.

Attention must also be paid to the possible effects of a changing price structure as money factor prices changes. So far we have assumed that a money factor price change is accompanied by proportionate changes in all other prices, so that price-cost relationship and relative income distribution are not seriously affected—in effect, we have neglected the effects of connected changes in the price structure. In a complete theory of income and employment these effects are potentially important enough to be taken into account. This is particularly true of changes in profits which result from price and income movements. For present purposes, we will simply point to the desirability of such elaborations of the theory, which, although they would not greatly alter our main conclusions, would make for quite significant amendments in detail.

INTEREST, INCOME, AND EMPLOYMENT

Let us now condense our preceding arguments concerning interest on money and the level of income and employment. In any initial situation of money and real income, it will be true either (1) that, with a given amount of money, the rate of interest will be so determined as to equate the demand for and

supply of investable funds, or (2) that, with given bank rate of interest, banks will add to or subtract from the money supply sufficiently to balance the demand for and supply of funds. In either event, saving may be greater or smaller than investment at the initial level of income, and this inequality may not be subject to correction by rates of interest which the banks can support. Where saving initially equals investment, money and real income will remain stable. Where they are unequal, the initial level of income is untenable, and money income will change until investment and saving become equal. Adjustments of money income to such a point of equilibrium may involve changes both in money prices and in employment. But of these, money price changes will not generally serve to overcome an inequality of saving and investment, since these are primarily determined as proportions of real income; as money amounts, S and I tend to respond together to price changes. Therefore, employment must move until saving and investment are equated by virtue of real income changes. We thus derive the idea of an equilibrium level of employment, such that saving out of the corresponding real income is equal to investment. This equilibrium employment may be high or low, depending upon the volume of investment demand current at the time.

In any event, it is largely independent of the price level with a fixed bank rate of interest and flexible money supply. For a given equilibrium level of employment, the equilibrium level of money income will differ according to the money price level. The equilibrium level of employment is determinate in any given situation. The corresponding levels of money income and prices are not so determinate but depend largely on institutional forces affecting the relative rigidity of money factor prices. The money price level does not generally affect the level of employment, except so far as speculation on price changes are involved, or except so far as the money supply from banks is inflexible and the price level thus influences the interest rate.

The outstanding conclusion, of course, concerns the tendency of the economy toward an equilibrium level of employment, possibly involving underemployment. This tendency results from the fact that employment moves until investment and saving become equal. Understanding of the principles under-

lying this conclusion is the key to the interpretation of transient or chronic unemployment, and of fluctuations in employment generated by fluctuations in investment. The argument is traced only in outline here, and should be pursued in much more detail. Detailed analysis of a period-by-period sequence of changes through time will reveal many phenomena not touched upon here, so that the preceding should be viewed only as a sketchy introduction to the analysis of the dynamics of income and employment.

In the general analysis of income distribution in Chapter 10, two main problems were outlined: (1) what determines the relative price per unit and the share of income received by each factor of production, and (2) what determines the level of employment of the factors of production, collectively and severally. We were there able to give a broad general answer to the first question by indicating that in a competitive economy with full employment of several productive factors, all of which were in limited supply, relative prices would be determined by technical substitution ratios among the factors. With all-round partial employment (should it occur for any reason) the same rule seemed to apply. To the second question, concerning employment, we did not develop an answer. Instead we had only indicated (1) that full employment would always ensue if competitive adjustments of money prices took place and if these adjustments did not cause offsetting adjustments in money income, and (2) that unemployment might ensue if an appropriate adjustment between money income and money prices did not take place, because either of money price rigidity or of a failure of income recipients as a whole to spend as much as they received. We did not, however, develop a theory of the relation of money income to money prices and, hence, of the determination of employment.

Let us now see what we have added to our knowledge on these two problems by our investigation of interest and capital. The most important addition has been with regard to the determination of the level of employment. The "factors of production" to be employed at any given time consist of a supply of labor, a supply of land, and a supply of existing capital goods. These are the productive resources which may be fully employed—so that all of them that seek employment at going

prices are used—or less than fully employed. In addition, there is a supply of investable funds, but as we have seen this does not qualify on the same grounds as a factor of production; and whether or not such funds are "fully employed" is not an issue, or necessarily a meaningful question. The significant question, therefore, concerns the level of employment of labor, land, and capital goods in the economy—the employment of available physical resources. Now we have seen that the employment of such resources does not, even under competitive conditions, automatically move to the full-employment level. Rather, employment moves to such a level that saving from the corresponding real income—*i.e.*, the portion of that income not spent by income recipients on consumption goods—is equal to investment—*i.e.*, the amount spent on capital goods plus loan-financed consumption. The level of employment thus depends primarily on the rate of real investment (which may be variable over time) and on the relation of saving (as defined) to real income. Employment cannot rise above the level where saving equals investment.

This, in turn, implies that the ratio of money prices to money income does not adjust to insure full employment. Movements in money prices generally will not alter the equilibrium level of employment except so far as they engender speculation and thus influence investment, or as they influence liquidity and the rate of interest. Money income will tend to adapt itself to any ruling level of money prices in such wise as to maintain the same equilibrium level of employment—a level determined by real investment and the relation of saving to real income. The rigidity of money prices is thus not a significant cause of unemployment, except so far as flexibility might bring speculative forces into play. There may, of course, be full employment, if investment is large enough to balance the saving from full-employment income. It is even possible that progressive inflation of money income and prices at full employment may result because of a current excess of investment over saving at a full-employment level. But it is also possible that there may be an equilibrium with involuntary unemployment, in which reductions of money prices will reduce money income but will not reduce unemploy-

ment.[19] A tentative explanation of the level of employment is thus developed, in terms of the determinants of real investment and of the relation of saving to real income. This relation determines in general the ratio of money income to money prices, and thus the level of employment. The absolute level of money income, on the other hand, is determinate only if movements in money prices are effectively explained.

The equilibrium level of employment just described may be influenced by the rate of interest on investable funds, as ordinarily set by the banking system. Employment may be somewhat increased or decreased by lowering or raising the interest rate, and by thus influencing saving and investment. But the bank rate operates theoretically within a range set by a positive theoretical minimum and some positive theoretical maximum, and practically within a narrower range. Moreover, neither saving nor investment may be particularly elastic to the interest rate within this range. Thus the rate of interest, even when supported by a reasonably flexible money supply from the banks, cannot ordinarily be used to alter seriously the character of a given employment equilibrium.

To the preceding it is necessary to add that employment "equilibrium" is not intrinsically a stable thing which tends to be perpetuated at a single level over time. It is an expression of the tendency of employment in response to any current rate of investment, given the relation of saving to real income and given the bank rate of interest. Since investment—and particularly its net investment component—moves or fluctuates over time, employment equilibrium is inclined to be a moving or fluctuating equilibrium. Consider the business cycle.

THE SHARE OF INCOME EARNED AS INTEREST

The share of the total income stream going to capital goods—for their initial purchase and periodic replacement—is largely an indirect payment of wages to labor and rents to land. This is because capital goods are essentially commodities produced with labor and land. This income stream reaches labor and land

[19] Unless via reductions in the interest rate, or by engendering speculation.

as they are used to produce capital goods for net additions or replacement. If there were no *cost* of capital goods other than the wage and rent payments directly or indirectly involved, all income under competitive conditions would go to wages and rents, as prices were driven by competition to the level of costs. A certain proportion of land and labor would be employed in indirect or roundabout fashion to produce capital goods, depending upon the character of production techniques, but all income would in any event go to wages and rents. There would in long-run equilibrium be no profits, and there would be no net or additional return on capital goods.

Any net or additional payment arising from the use of capital goods must be in the nature of *interest*, and this emerges as a share of income because funds must be acquired for investment in capital goods, and because there is an interest charge or cost which must be paid to secure funds for investment. Interest must be paid to acquire funds for investment and to keep them invested; this payment is included in costs of production; prices tend to be adjusted under competition to equal total costs inclusive of interest; and a part of the income stream thus flows to the suppliers of investable funds. The total share of income which flows as interest will thus evidently depend upon the supply price of investable funds and on the total volume of funds invested. What the supply price of investable funds is or would be under various circumstances can be analyzed, but the primarily relevant circumstance is that of the modern economy with a highly developed central banking system. Here and today, the rate of interest is an arbitrary rate set by the banking system and supported by a flexible supply of money. It will not be lower than some positive minimum at which the demand for liquidity would be unlimited, or higher than a maximum at which the demand for securities would be greater than the banking system can meet, and thus it is limited ultimately by the liquidity preference attitudes of the populace. But between these limits it tends to be an arbitrary rate set by banks in line with over-all economic and political policy.

In recent years, the bank rate (exclusive of risk premiums) shows a tendency to be relatively stable at levels between ½ percent and 2 percent per annum. The maintenance of some such

rate, possibly with periodic adjustments for varying levels of employment, essentially means that the supply of investable funds tends to be very elastic or even perfectly elastic at the bank rate of interest. That is, funds will be supplied in any amount ordinarily demanded at the rate the bank is supporting, the bank standing ready to meet any demand not met by nonbank sources or to absorb any supply not taken by nonbank demands. Funds thus tend to be in very elastic supply, and investors will tend to earn, net of risk premiums paid to counterbalance defaults on principal, the same bank rate of interest, more or less regardless of the level of total investment. Given this relatively fixed price of funds—say 2 percent—all invested money will tend to earn a net 2 cents per dollar per annum. The total *amount* of income going as interest will then depend mainly on the amount of funds invested, and the *share* of income will depend upon this and upon the ratio of total accumulated investment to total income.

Investment of funds occurs primarily in capital goods, and secondarily in consumer loans. Either will yield a rate of interest and provide the supplier of funds with a share of income while his funds are invested. The quantity of investment in consumer loans is not easy to appraise analytically, but it would appear to vary roughly with the level of income. Some relatively stable though small proportion of total income thus flows as interest on consumer loans. The quantity of investment in capital goods tends to reach a limit in any given state of techniques, products, population, etc.—and within such a state it will vary somewhat with income and employment. Currently, the quantity of investment is very large, and the share of business income paid as interest is sizable. With progressive changes in techniques, population, and so forth, total investment in capital goods has tended historically to increase. But because of corresponding increases in total income, resulting from increases in efficiency and in amounts of other factors, the ratio of accumulated investment to income has not increased indefinitely, and apparently tends after a point to become relatively stable. Although the continually increasing use of capital goods has raised the amount of income paid to interest, it has been matched in part at least by corresponding increases in wages and rents, so that the share

of income going to interest has not increased in proportion. Summarizing, it is true, with a relatively fixed interest rate, that the share of income going to interest at any current time is substantial but limited by the limitation on the demand for capital goods and consumer loans, and that the increase of the interest proportion of income as accumulated investment increases over time is retarded or entirely checked by the associated increase in total income.[20]

In any event, a certain share of income in a capitalist economy flows as interest paid to investors for the investment of their funds, and this payment tends on the average to be at a rate on invested funds corresponding to the bank rate of interest. Does this mean that every investor receives every month a rate corresponding to the current market or bank rate on the current valuation of his investment? This would be true if the amounts of capital goods would be currently adjusted without lag to changing interest rates, and if all investment contracts and loans were continually rewritten from day to day as necessary to adjust for changing money market conditions.

In fact, of course, neither of these conditions is fully observed, and as a consequence the amounts "earned" by capital goods may vary from the current market rate of interest on original investment value, and the amounts received by investors as interest payments may vary from the current "earnings" on their investments. Investments in capital goods which have been made to "earn" a 2-percent rate of interest will continue to yield at this rate even though the market rate of interest later drops to 1 or rises to 3 percent, until the amount of capital goods can be adjusted. Thus the enterprise may currently realize from its capital goods an earning different from the current market rate, and this discrepancy can be corrected only in the long run. Further, the money earning on capital goods may vary in response to price and income variations in the economy, so that the investment yields more or less than was anticipated at the time it was made. Thus a 100-percent increase in the price level

[20] Additional interest returns are of course earned by those who buy non-wasting assets (land) at discounted present value of future rents. We must add these to interest returns from capital goods and consumer loans to arrive at the full total for the economy.

might double the imputed money earning of a $5000 original investment in a machine—might double its annual earnings (say from $150 to $300) actually raising its "value" to $10,000. If, in addition, the investor who supplied the funds holds a $5000 3-percent bond, he will continue to receive annual interest of $150, and the enterprise will receive the balance of current interest as a "profit." The current flow of earnings on capital goods may thus deviate from the current rate of interest on original investment. Further, the amounts received by investors, and especially by creditors on fixed-interest contract, may deviate from current earnings. These discrepancies arise respectively from deviations from a previous equilibrium adjustment, which influence the income flow, and from fixed investment contracts, which put the enterprise in a position to pay to investors more or less than current economic interest and to absorb the difference. Subject to such qualifications, our generalization concerning the interest share in income is substantially accurate.

Who receives the interest share? This is more than a needlessly obvious question, because funds are supplied by banks as well as by individuals. The interest share of income, then, flows in the first place to individuals who have supplied funds for investment, or to their descendants who inherit their accumulations. These accumulations originally tend to result from savings out of income, although the income may be variously earned, and, if it is large, saving may not result in any perceptible degree of privation on the part of the saver. Such interest is a reward to individuals for accumulating and making available funds for investment; it is "needed," if not to induce saving, at any rate to induce investors to part from liquidity. In the second place, however, an interest share of income also flows to banks, in return for the supply of funds which they create by expanding their credit. Under modern conditions, a bank is a special sort of private business enterprise, licensed and controlled by the government, one of whose functions is to create credit. Banks are essentially franchised to "manufacture" money, albeit under strict limitations, and to make it available in return for an interest payment. A part of interest is thus a payment to investors in banking enterprises, received in connection with the performance of this special function. It is worth re-emphasizing that the

government, through its central bank organization, undertakes to regulate this rate of earning, and thus also the rate of interest on individual accumulations of funds.

Three additional matters now deserve comment in connection with a general discussion of the interest rate: risk premiums, monopoly and monopsony in the market for funds and for capital goods, and the place of governmental debt in the interest picture.

RISK PREMIUMS AND INTEREST

It has been emphasized at several points in the preceding discussion that the interest rate to which we refer is the net rate paid for the loan of funds, exclusive of any premium allowed to cover the risk of loss. This is the "pure" interest rate, or payment necessary to secure liquid balances in return for securities even though the lender or investor is "100 percent certain" of receiving all contracted or expected interest payments for the duration of the investment, and also of receiving the entire original principal amount at the expiration of the investment period. It is the rate payable if the investor is "fully guaranteed" against any risk of loss. "Certainty" may be an inappropriate term for use in referring to the uncertain world, but the idea is given meaning if we define a "riskless" loan as one for which lenders regard the risk of loss as sufficiently improbable that they disregard it entirely. The standard example for a riskless security is a short-term note or bond issued by a financially strong and responsible government and backed by the taxing power of the state. Here the possibility of default on principal or interest is negligible, and the rate which the government must pay to borrow funds approximates very closely a "pure" interest rate. In effect, the government has to pay practically nothing as an insurance against risk—what it does pay is therefore entirely interest. In recent years, this rate has ordinarily stood, under the influence of a liberal central bank policy, at 1 percent or less.

Such a "pure" interest rate must be paid on all loans (and under fully competitive conditions should be equal for all of them) and it is to this rate which we have been referring at length above. Most loans or investments, however, are not fully

guaranteed against risk. For private or nongovernmental loans or investments in general, there is a recognizable possibility of loss—a calculable risk that the enterprise or individual using the funds will not be able to meet interest payments or to return part or all of the principal. Business investments in capital goods, for example, are made on the basis of anticipations of future earnings. If these anticipations are badly disappointed (and they may be, because the future is never certain) the enterprise may be unable to recover from past investments as much funds as went into them, and may be unable to repay the principal of loans or to sustain the value of more permanent investments. How does this possibility affect the payments received by investors?

It logically need not affect them at all, if on the one hand (1) investments were always made with such conservatism or consistent abhorrence of the chance of loss that no net risk remained, or (2) if the firms or individuals acquiring funds always "insured" investors against loss, by paying some sort of an insurance premium to a casualty company which in return therefor assumed all risks connected with the investment. It does, in fact, affect the payments received by investors because investments involving risk are made and because the risk is not ordinarily insured by a third party. In effect, investors are asked to assume the risks and to "insure themselves" against loss, and in return for this they are paid, in addition to pure interest, a "risk premium" to compensate for the calculated probability of loss. Thus the stated interest payment on a given private security is ordinarily a compound of pure interest and risk premium, which together make up the gross so-called "interest payment." Thus if riskless government securities bear 2 percent interest, and a corporate bond issued to finance a new venture bears 5 percent, it may be inferred that the investors in the latter security are being paid 2 percent pure interest, plus a 3-percent premium against the risk of partial or total loss. Viewing the security markets as a whole, there will be a great variety of gross "interest" yields, representing different degrees of risk and correspondingly different premiums offered to compensate for them.

This accounts fairly well for the variations among the yields on securities where a given payment is contracted, as on bonds, notes, and other explicit loans. At least a part of investment in business enterprises, however, is in the nature of "equities"—*i.e.*, in return for ownership certificates or shares of stock which are guaranteed no specific payment but stand to receive, as dividends or gain, the title to any residual earning left after the enterprise has met all contractual "interest" and all other costs. Such investments, in turn, bear the largest share of the total risk in an enterprise—the shareholders of a corporation, for example, in effect pledge the entire principal of their investment against the fulfillment of all obligations to creditors who invest in the business by loaning to it.

Although the exact "risk premium" allowed to such shareholders is not stated or contracted for, however, it may be imputed in much the same way as if it were. That is, such investments are made upon the basis of an expectation of a pure interest return plus a sufficient risk premium, if the business "succeeds," to offset the probable loss which will be incurred if it "fails." Shareholders may pay $100 per share of common stock, for example, upon the expectation of a $7-per-year dividend. Of this only $2, for example, may be pure interest. The additional $5 per year, or 5 percent, is required to offset what amounts to a one-in-twenty chance, from year to year, that the business may go bankrupt and realize nothing for the shareholders. Thus we may include ownership or equity securities along with contractual loans as bearing a pure interest rate plus, implicitly, a risk premium. We will return to this "implicit" premium below.

Of what importance to the economy are these risk premiums which must be paid on investable funds in order to finance the purchase of capital goods, consumer loans, and so forth? For the economy as a whole, risk premiums should not constitute a net share of income, so far as risks are on the average accurately anticipated. Actual losses should occur frequently enough on the average to offset the risk premiums paid on investments which succeed, so that the aggregated net "earning" from the undertaking of risk should tend toward zero. If twenty $100 investments pay an extra 5 percent against the possibility of loss, and

the investors are thus enriched by $100, a twenty-first should be lost entirely, thus balancing the account at zero. The individual investor may gain or lose by undertaking risk; the aggregate of investors break even. Such a generalization, of course, would tend to hold only on the average and in the long run. With fluctuating income in the economy, over-all gains or over-all losses may occur over considerable time intervals, as anticipations are found to have been, for the time being, too conservative or too sanguine. The only possibility of a consistent income share deriving from risk premiums would arise from a persistent tendency on the part of investors to overrate or to underrate the risk of investments—from a persistent error in anticipations. In this event, a net income share, positive or negative, could arise for the economy as a whole. Such a share might also be created by the monopolization of funds, together with a deliberate overcharge for risk, although this should probably be accounted as a monopoly rather than a risk-bearing gain. Such possibilities may be more appropriately explored in more specialized works than the present one. Neglecting them, the general conclusion is that risk bearing tends to be rewarded by no net share of total income.

The incidence of risk, however, will have a definite effect on the rate of investment in capital goods and on the demand for investable funds. The firm acquiring funds for investment must allow for the added "risk" charge as a cost of the investment, and will regulate its investment accordingly. The firm borrowing funds, for example, must pay a gross rate of 6 percent to bondholders instead of the current pure interest rate of 2 percent. "Money" costs it 6 percent instead of 2; this raises the cost of capital goods, and investment is somewhat restricted.

The firm using the funds of its owners will similarly restrict investment to compensate for the possibility of loss; it will do so by choosing some rate of investment *which seems as likely to yield above the current pure interest rate as it is likely to fall below.* There is no essential difference between the two cases. The owners with their own money regulate investment so that the "mean expectancy" is for a yield equal to this pure interest rate—so that the possibility of extra gains above this rate just counterbalances the possibility of losses, or of returns less than

pure interest plus principal. Investment is set low enough to allow for the possibility of an extra gain sufficient to offset possible losses. The owners using borrowed money follow the same calculation (neglecting differences between the creditors' and their own appraisals of risk); investment is similarly restricted. But since the creditors undertake the risk of loss without automatically sharing in extra gains if they are made, the risk premium is explicitly added to the interest rate on loans. Explicit risk premiums thus arise only on loan capital, but risk is a "cost" and thus has a restrictive effect on investment of all types, whether by owners or creditors. Its total burden for the economy is the aggregate of monetary losses on all unsuccessful ventures—this loss tends on the average to be borne as a cost by the ones that succeed, since they must earn "enough extra" to offset this loss and thus still make it attractive for people to supply investable funds to enterprise.[21]

It should not be inferred, of course, that any and all earnings of the shareholders of enterprises are therefore "normal" returns for risk. These may include as well genuine "excess profits," from monopoly or other sources. But the appraisal of returns on individual securities must allow for risk premiums sufficient on the average to offset the burden of losses for investments as a whole.

MONOPOLY AND MONOPSONY AS THEY AFFECT CAPITAL GOODS AND INVESTABLE FUNDS

We have pursued our discussion of interest and investment so far on the assumption that there is for the whole economy a single "competitive" market for funds, unaffected by monopoly power on the part of suppliers or monopsony power on the part of borrowers. Funds would thus be secured for investment throughout the economy at the same basic pure rate of interest. To what extent do these assumptions lead us into inaccurate

[21] The present treatment of risk is extremely simplified and omits certain significant phenomena arising from uncertainty. For suggestions of a more advanced treatment, see A. G. Hart, "Risk, Uncertainty, and the Unprofitability of Compounding Probabilities," *Readings in the Theory of Income Distribution*, Chap. 28.

conclusions concerning interest and investment? A full analysis of banking operations and of money and security markets, which would be essential to the answering of this question, lies outside the scope of this volume. Nevertheless it may be useful to comment briefly on the potential importance of monopoly and monopsony in the market for investable funds.

The meaning of a "competitive" price or interest rate for funds is rather special, since it is essentially a rate dictated by the central banking system according to broad principles of policy. The central bank endeavors in general to support an interest rate and to provide a perfectly elastic supply of funds at this rate throughout the economy. All other suppliers of funds are in a general way in competition with this rate and tend to be forced toward it so far as funds originating with the central bank effectively compete with them in supplying particular borrowers. The competitive rate is thus the rate supported by the central banking system. Monopoly power in supplying funds would consist of the ability to charge a higher than competitive rate; monopsony power would consist of the ability either to secure a lower than competitive rate or to negate monopoly power and secure a competitive rate in spite of monopoly. What we are basically interested in, then, is deviations of particular interest rates throughout the economy from the rate supported by the central bank.

Monopolistic deviations may stem potentially from two facts. First, the market for funds is not a single unified market accessible to all borrowers. Rather it is made up of one or more central markets accessible to large borrowers or to securities issued by these borrowers, and of a large number of small submarkets each with a group of small borrowers who because of geographical and institutional barriers do not have full or direct access to other submarkets or to the central money markets. Second, the various submarkets are not directly supplied by the central bank but are only indirectly influenced by it. The specific suppliers in submarkets may be a few commercial banks or other lending institutions which have some monopoly power over the borrowers in these markets and can exact what is implicitly a higher than competitive rate of interest. For such local markets the central bank may attempt to influence the interest rate by

supplying the commercial banks with funds at the competitive rate. But these and other lending institutions may exploit their local monopoly position and charge a higher rate to borrowers.

For many small borrowers in local markets—especially for businesses too small to establish a "credit" or risk rating in the central markets and for consumers who wish to borrow to finance purchases—there may thus be interest rates somewhat in excess of the "competitive" level. Such monopolistic pricing of funds is restrictive of investment (or may tend to divert investment opportunities to large business firms) and is also a potential source of extra earnings to monopolistic suppliers of funds.

Large borrowers, on the other hand, are not limited to a single submarket and generally have access to the central money markets of New York. For them, the danger of monopolistic exploitation in the supply of funds is decidedly less, both because they have access to a variety of suppliers, and because in many cases they are big enough that their monopsony buying power serves as a protection against exploitation. Nevertheless it may be noted that in the organization of investing institutions and investment banking in the central markets there are possibilities for the monopolistic control of very large amounts of funds, and that automatic adherence to the "competitive" rate is not a foregone conclusion. Viewing the picture as a whole, we must take some account of deviations of the interest charge from the competitive level on specific loans or classes of loans. Such deviations are probably not persistent enough or large enough to require any serious alteration in our general conclusions. But a detailed study of money and interest would reveal a great many matters which we have overlooked.

The potentially most important avenue by which monopoly may affect investment is, in fact, not in the pricing of funds but in the pricing of the capital goods in which funds are invested. Such capital goods are produced commodities, and they are often produced under monopolistic or oligopolistic conditions and sold at considerably higher than competitive prices. Now the disposition of enterprises to invest is based on the economy of substituting capital-goods services for other services, and the extent to which this substitution is carried depends on the relative

prices of the alternative services. Inclusion of monopoly profits in the prices of capital goods raises the ratio of the prices of their services to that of others and tends to limit investment. In fact, it may do so to a much greater extent than moderate increases in the interest rate. A 20-percent increase in the price of a $100,000 machine lasting ten years adds $2000 per year to the cost of using the machine; a 100-percent increase in a 2-percent interest rate would be required to have the same effect. So far as monopolistic pricing of capital goods occurs—and many of the basic "producer" goods industries have highly concentrated market structures—it tends to restrict the total amount and the current rate of investment. In this way, monopoly may make a specific addition to the tendency to unemployment.

Specific evaluation of the character of capital-goods pricing, however, must also take into account the possibility of monopsony and of bilateral monopoly situations. It may be true that the purchasers of or investors in capital goods have a monopsony power, in that they can depress the price of these goods by restricting their purchases of them. Such monopsony power may tend to characterize a number of industries buying capital goods, since such industries are often concentrated in structure and their firms thus able to influence their buying prices. What influence does such monopsony power tend to have in the absence of counterbalancing monopoly power? When we investigated the idea of monopsony, and of related oligopsony, in Chapter 7, we saw that the monopsonist balances *marginal outlay* on the monopsonized good against his "demand" for the good. This leads generally to a depression of the buying price of the good, but the effect on total amount purchased depends on the conditions of supply for the monopsonized good. If there is a positively sloped supply curve, purchases are restricted; if there is a negatively sloped supply curve, purchases are increased above the competitive level.

This general principle may be extended to cover monopsony purchase of capital goods. If the purchaser faces a positively sloped supply curve for the capital good he buys (Fig. 37, p. 225) he will tend to restrict the degree to which he substitutes the capital good for other services, thus restricting investment

for any given output,[22] and to restrict his output, so far as increasing marginal outlay with increasing output contributes to a more rapid rise of his own marginal cost of production. For both of these reasons, total investment would tend to be restricted by monopsony if the supply curve of the monopsonized good were positively sloped. If, on the other hand, there were a negatively sloped supply, so that the price of the good fell with increased output, investment would tend to be enhanced. In the case of capital goods, the second situation seems likely to be just as common as the first, and the supposition that monopsony has a considerable effect on the rate of investment in capital goods is not well supported by known facts.

There are a number of situations, however, where they are few buyers and few sellers of capital goods—in effect bilateral oligopoly. In this case it seems safe to apply our generalizations from Chapter 7 and suggest that the net effect on the rate of investment is theoretically indeterminate.[23]

Summarizing on departures from competition in the determination of interest and investment, we must make a number of detailed modifications of the "competitive" analysis if we are to come close to the facts of individual cases. The only serious modifications, however, concern monopoly and monopsony in the pricing of capital goods, and the effects of these conditions are readily taken into account. Otherwise, the conclusions of the competitive analysis stand without serious over-all modification. We will return to the general effects of monopoly and monopsony pricing of productive factors again in connection with the discussion of labor and land.

We have previously referred to the fact that there may be not simply a pure rate of interest but a family of pure rates for investments over various time periods. The long-term interest rate, which might be paid, for example, on bonds of 40-year maturity, may differ from that on one-year loans, and so forth. This discrepancy among rates is essentially a reflection of anticipations on the part of suppliers and borrowers of funds of an

[22] This is because he now balances the marginal outlay on the monopsonized good against the prices of non-monopsonized goods in determining the most profitable rate of investment.

[23] See the further discussion of bilateral monopoly in Chap. 13.

increase or decrease in the interest rate over time. Thus if the current rate on short-term loans is 2 percent, the long-term rate will tend to be higher or lower, depending upon whether it is expected that future short-term rates will rise or fall. The lender will not tie his money up at 2 percent for 10 years if he feels that the *average* short-term rate over the 10 years will be 3 percent. The borrower will not agree to pay 2 percent for 10 years if he feels that the average short-term rate over that time will be 1 percent. The long-term rate for any specific time interval will tend to represent an average (strictly a kind of geometric mean) of expected short-term rates over the interval in question. This expectation refers to an equilibrium *marginal* expectation effective for pricing in the money market. Thus at any time there may be a family of interest rates for loans of various maturities, reflecting the prevailing state of anticipations regarding future short-term interest rates. Even so, there would not be different rates for different time intervals if the central bank undertook to support the same rate for all types of securities. The discrepancies can arise because the banking system centers *principally* upon short-term loans, and allows the expectations of private investors some free play in the determination of longer term rates. The effects of a family of interest rates must be taken into account in any detailed appraisal of investment.

THE GOVERNMENT DEBT, INVESTMENT, AND INTEREST

We have so far referred to the use of investable funds entirely in terms of private or nongovernmental investment, undertaken by business enterprises for a profit or by individuals to increase their current purchases of consumer goods. This is the appropriate emphasis in a capitalist economy, where the bulk of investment activity is privately conducted and based on motives of individual advantage. Nevertheless, the governments under which capitalist economies operate may also borrow funds, potentially in very large amounts, to finance routine or emergency expenditures. Government debts tend to increase most in times of war, when it is politically more expedient to finance armament expenditures by borrowing than by taxes. Governmental

agencies may also borrow to finance peacetime expenditures, however, especially those on public works, which are generally capital investments which private enterprise may be reluctant to undertake. Under the pressure of prolonged wars, the government debt may easily become comparable in magnitude to the total private debt, and its accumulation, maintenance, and possible repayment may have considerable effects on the level of activity in the private economy.

Additions to the government debt—current government expenditures financed by a deficit—constitute an addition to the current flow of "investment" in the economy. When the government borrows (from individuals or banks or both) and spends 10 billion dollars, this is a gross addition of 10 billion to expenditure in the economy; so far as the government borrowing operation has not caused any curtailment of private investment or consumption, it is a net addition to expenditure. If the government is maintaining a fixed interest rate through the central bank, and if it does nothing to encourage saving, most of the addition should be net. Government deficit-financed spending, with corresponding *additions* to government debt, thus tends to create income in much the same fashion as private investment. Income for a period will tend to move to the level where saving from it is equal to private investment of the period plus the government deficit of the period. The possibility of using governmental deficit finance to create larger income suggests its use as a means of lessening unemployment in times of depression or chronic stagnation, and it was so used in the period from 1933 to 1941. Wartime deficit-financed spending has the same general income-generating effect, but often the wartime deficits are so large as to keep income rising (with investment above saving) after full employment is reached, and thus to propagate considerable price inflation. Since at least a part of additions to the government debt will ordinarily go to the banks, the rise in income which it generates will be matched by some corresponding rise in the amount of money in circulation, and possibly by increased liquidity relative to income.[24]

[24] For a discussion of the issues raised above, see A. H. Hansen, *Economic Policy and Full Employment.*

Additions to the government debt thus tend to have the same general effect on current income as does private investment. What of the effect of the maintenance or "servicing" of the debt through interest payments? Here we must draw one distinction. So far as the government has invested in revenue-producing projects, such as hydroelectric plants or toll bridges, its operation is very much like that of a private enterprise, in that it collects from users of the service and disburses interest payments along with other costs. So far as its investments are not revenue-producing, however, either because the expenditure results in no lasting assets—as in the case of munitions—or because it is not expedient to collect directly for the service—as in the case of many educational or recreational projects—then the debt must be serviced from tax revenues. In this case, a share of income is diverted from taxpayers and paid as interest to government bondholders. Since World War II, an important portion of the total "interest" share in income is so collected and paid to bondholding individuals and banks. The burden of this interest payment is borne by people not necessarily in proportion to benefits accruing from the original investment, but in such proportions as are determined by the principles of taxation in effect. The "cost" is borne in a manner determined by governmental decision rather than by a free market. Large government debts thus tend to introduce an important inflexible element into the pattern of income distribution, and to make taxation an important instrument in redistributing income in accordance with a rather fixed pattern.

The repayment of the principal of a government debt is by no means inevitable, since it may be "refunded" indefinitely over the future. When repayments are undertaken, they may tend in general to have the reverse effect from additions to debt. That is, tax revenues which in part at least tend to reduce consumption spending are used to redeem bonds as they mature, and the former bondholders (including banks) will tend not to spend most of their redemption payments on consumption. Thus debt repayment tends in general to be a deflationary or income-destroying operation and is best undertaken in periods of superabundant income. The preceding few remarks, however,

hardly constitute a thorough review of governmental fiscal operations, and for this we must refer the student to more specialized works.

SUPPLEMENTARY READINGS

IRVING FISHER, *The Theory of Interest*, Chaps. 1-5.

JOHN MAYNARD KEYNES, *The General Theory of Employment, Interest, and Money*, Chaps. 8-10, 13-21.

ALVIN H. HANSEN, *Fiscal Policy and Business Cycles*, Chaps. 1-3, 11-14.

GOTTFRIED VON HABERLER, *Prosperity and Depression* (2d ed.), Geneva: League of Nations, 1939, Chap. 8.

ERIK LUNDBERG, *Studies in the Theory of Economic Expansion*, London: King, 1937.

HOWARD S. ELLIS, "Monetary Policy and Investment," *Readings in Business Cycle Theory*, Philadelphia: The Blakiston Company, 1944, Chap. 20.

FRIEDRICH A. LUTZ, "The Structure of Interest Rates," *Readings in the Theory of Income Distribution*, Philadelphia: The Blakiston Company, 1946, Chap. 26.

FRANK H. KNIGHT, "Capital and Interest," *ibid.*, Chap. 21.

CHAPTER 13

THE THEORY OF FACTOR PRICING AS APPLIED TO RENTS AND WAGES

A remaining task in this analysis of pricing and income is to consider somewhat more specifically the determination of the labor share in income, wages, and the share paid to resource owners as rents. Under modern conditions of bilateral monopoly bargaining in most labor markets, not too much can be learned by the a priori analytical methods to which this volume is devoted. For knowledge of wage determination, we must at present rest heavily upon the largely unsystematized observation of individual cases. But it may be useful to consider briefly some general aspects of the problem of wage determination and to examine at the same time the theory of rent.

Let us see how the analysis before us is related to that contained in preceding chapters. The volume to this point has been concerned with the following matters:

1. The general determination of commodity prices and of their relation to factor prices (costs of production), with some given flow of money income (Chaps. 4 to 9).
2. The adjustment of factor prices and employment of factors, relative to any given level of money income, under conditions of pure competition (Chap. 10). For this purpose, each factor was viewed as some inanimate basic commodity with given supply conditions.

3. The determination of the use of capital goods in connection with labor and land (Chap. 11).
4. The determination of the level of money income and employment relative to any given level of factor prices (including the interest rate as given), and also of the general relation of factor prices, money income, and employment where all may vary (Chap. 12).

In the preceding chapter it appeared that with any *given* level of wages, rent, and interest, the relation of saving to investment could be of several different sorts and correspondingly have several different results. One possibility is that there may be a tendency to "overemployment." That is, investment may tend to exceed saving persistently at the given level of factor prices even when full employment is reached, so that the aggregate demand price for goods continually runs ahead of the aggregate supply price. Then there would be a progressively expanding money income, a tendency to rising factor and commodity prices, and a dynamic process of expansion, "at full employment," which would not be easily stabilized. A second possibility is that of a tendency to "underemployment," or involuntary unemployment. That is, investment may exceed saving at a full-employment real income and equal saving only if employment and real income fall below this level. Then not all factors which wish to be can be employed at any going level of money factor prices. This situation is also dynamically unstable unless money factor prices (or at any rate one major price, like wages) are pegged or rigid, and thus become a pivot point for a stable underemployment equilibrium. In the absence of such factor-price rigidity, there would tend to be a dynamic process of contraction of money income and of falling prices, with some noneliminable margin of involuntary unemployment.

In either of these two cases, there is no certain stable equilibrium relation of factor prices to commodity prices, or of one factor price to another. At the best there are virtual equilibrium tendencies (as of a full-employment equilibrium) which may be approximated, though poorly, in the process of dynamic change. We thus have no dependable theory of pricing, for fac-

tors or commodities, to apply to situations of dynamic instability of income and employment.

A third possible result, however, is that there will be stable equilibrium for the economy just at the level of full employment. That is, investment may equal saving when all factors which wish to work at some going level of factor prices are employed, and the aggregate demand price for goods will equal their aggregate supply price from period to period through time. Any initial level of money income is then self-sustaining at a constant level through time. A flexible interest rate may create a certain limited range within which such an equilibrium may occur, but an investment demand which will put the economy within this range is more or less a happy coincidence. Stability at full employment is thus not a major probability for a free-enterprise economy, which is likely to be involved in dynamic instability on one side or the other of this potential equilibrium range.

Should a stable full-employment equilibrium be attained, however, an equilibrium adjustment of commodity to factor prices and of factor prices to each other is possible, and then certain definite generalizations concerning price relationships can be applied. This is in contrast to pricing under conditions of dynamic instability of income and employment, where the determination of price relationships is quite uncertain.

As we look further into factor pricing, by considering in more detail the character of labor and land, we must therefore recognize that two orders of theoretical explanation of factor pricing are possible. First, there is a "stationary equilibrium" theory of factor pricing, appropriate to the third situation mentioned above, which in effect traces out the adjustment of factor prices to a given constant level of money income. This is essentially a theory of factor pricing under stable full-employment conditions. Second, there is potentially a "dynamic" theory of factor pricing, which should explain the serial behavior through time of factor prices as they interact with an unstable and moving money income. Although economic theory is fairly adequately developed to handle the first sort of situation, it is as yet hardly adequate to deal with the second. Most of that which can be set forth a priori concerning the determination of wages and

rents thus concerns what would tend to happen to them in a stationary equilibrium at full employment.

It will be convenient hereafter to deal first and principally with the theory of wages and rents as of such a full-employment equilibrium, and then to offer some comments on dynamic process adjustments and their general effects. This is not an entirely satisfactory procedure, but the current limitations of developed economic theory force us to it. Our initial concern will thus be with what *would* happen to wages and rents if there were a given constant flow of money income and if factor prices could work toward a final equilibrium relative to this flow, without at the same time disturbing it.

GENERAL CONSIDERATIONS AFFECTING WAGES AND RENTS

The basic analysis of wage and rent determination under competitive conditions is contained in Chapter 10, where labor and land were provisionally viewed as basic inanimate commodities with given conditions of supply. There we saw that, for any such pair of factors, equilibrium prices at full employment would depend jointly upon the technical substitution relations between the factors as employed in production and upon their respective conditions of supply in money or real terms. Given these determinants, an equilibrium employment and an equilibrium relative price for each factor would be found. The student is referred to Chapter 10 in its entirety for review on these points.

The elaborations of the equilibrium theory of wages and rents from this point will proceed basically by: (1) taking account of the peculiar conditions of supply of labor and of land, as they would exist even in competitive markets; (2) recognizing the effect of nonhomogeneity in the supply of either labor or land, and of the introduction of a series of related submarkets for different subparts of the labor supply or of the land supply; (3) recognizing the impact upon wage and rent determination of monopolistic conditions in the markets in which commodity outputs are sold; and (4) recognizing the effect of monopsony in the buying of labor and land, and of monopoly in their sale.

The conditions of supply for labor under a system of free

contract and without slavery are basically affected by the fact that labor is a human service sold by human beings performing it. The supply conditions are thus fundamentally a product of human psychology—of the interplay of the need for a livelihood, the desire to work, and the complementary "real cost" of effort and desire for leisure. The actual supply conditions in a labor market will nevertheless differ greatly between situations where labor is unorganized and each individual laborer offers his services in a purely competitive selling market and situations where labor is organized and is offered under monopolistic or quasi-monopolistic selling conditions by a labor group or organization. In the first case, supply depends upon individual psychology relative to work and leisure. In the second, it depends upon the character of group organization, of organization leadership and internal politics, and of group psychology. For contemporary purposes, the relevant reference is to labor supply by organizations. It may nevertheless be useful to refer briefly to the character of labor supply under competitive selling conditions.

It has generally been thought (1) that a competitive labor supply would tend to have some positive elasticity relative to the wage rate, and (2) that it would tend to be calculated basically in terms of the real wage rates offered to labor. The amount of labor offered by a large group of independent laborers would increase with an increase in the real wage rate and decline with a fall in this rate, because higher rewards (in terms of the purchasing power of money wages) will induce more people to work and employed people to work longer hours. Presumptively, the marginal offer of labor should balance the marginal disutility of added effort against the utility of the real wage received. The supply curve should run fundamentally in real terms. A derived supply curve relating money wages to the supply of labor would thus respond directly to changes in the commodity price level, shifting outward with any decline in commodity prices and backward with any rise therein.

With such a condition of supply for labor in a single purely competitive labor market, and with a given constant flow of money income, full employment would be readily attained at a real price where the total demand for labor just equaled the

number of labor hours offered for employment at that real wage. The particular amount of labor employed in such an equilibrium would depend upon the character of labor supply, as influenced by individual preferences as between work and leisure, and upon the real wage offered by enterprise, as determined by the productivity of all factors, the substitution relation between labor and other factors, and the relative supplies of various factors. Since labor is not actually supplied under purely competitive conditions, there is little point in pursuing this analysis in detail. We will refer to labor supply under monopolistic conditions at a later point.

The conditions of supply for land and resources are basically different from those for labor. Land is essentially an inanimate basic commodity, available in nature without any real cost of production in terms of human effort. The reward paid to land results from its scarcity—there is naturally little enough of it relative to labor that its marginal rate of substitution for labor gives it a finite price. The reward under a system of private property in land is a reward to passive ownership. It is paid to whoever owns the land because it is scarce and hence valuable, but not in return for any real effort purchased from the owner.[1] Because the services of land are available without human effort or real cost, it is commonly supposed that under competitive conditions they will be available in fixed supply regardless of their price. That is, there would be a perfectly inelastic supply schedule for land services equal to the total amount available, and this would not vary in response to variations in land price from zero upward. Any owner would be unable to influence the rents on his land, and since the marginal cost of supplying it is zero, would offer the entire amount for any price the market would offer.

[1] In this connection it should be emphasized that capital goods used in connection with land and resources (like mine shafts in coal mines or plows and fertilizer on farms) *are* capital goods and realize a distinct income share, in the form of depreciation and interest, which is not rent. It should also be emphasized that although no human effort is spent in making available the basic services of land, any owner may sell land to another, so that current holders may have transferred their hard-won cash balances to others in order to gain rights to the incomes from land. This does not erase the fact that land rents are basically not earned by human effort.

Under purely competitive market conditions in the sale of land, therefore, its price, rent, would always tend to fall enough to secure its full employment (unless corresponding declines in money income prevented this). Also, rent would tend to arrive at a level relative to wages and interest determined by the marginal rate of substitution of land for labor and capital at the point where land was fully employed. Rent thus would rest essentially on the scarcity of land. The less there was of it relative to other factors, in any given state of techniques, the higher the ratio of rent to other factor prices.

Earlier economists were much concerned with the determination of the ratio of wages to rents, and this was an appropriate emphasis in a primarily agrarian economy where agricultural land was the principal productive instrument other than labor. Their theories of rent contained the basic observations made just above, and pointed especially to the significant fact what with given techniques and a fixed supply of land, a rising population would mean lower wage rates and higher rents. (This would increase the marginal rate of substitution of labor for land and reduce the ratio of wages to rents.) Such a concern is still appropriate, although changing techniques have lessened the relative importance of land and enhanced that of labor and capital goods, so that in spite of increasing population the rent share of income has not increased progressively. It will be well to recognize, however, that monopolistic selling markets for the services of land, resulting from concentration of its ownership, may result in artificial as well as natural scarcity, and in corresponding deviations in the conditions of supply for land. We will return to this matter below.

We may also note at this point a certain divergence in the meaning of "real cost." Land, as noted, is available in nature without real cost for its production. ("Made land," like reclaimed swampland, is principally "a capital good.") Nevertheless, an increment in the amount of land used in producing a commodity may be counted as an increment to the "real cost" of producing the commodity, if in the second case we mean by "real cost" any increment in the amount of resources used, regardless of whether or not human effort is expended in making

a resource available. Real cost is appropriately used in the second sense in arguing about the optimal allocation of all resources among the production of various commodities.

DISPARATE SUBMARKETS FOR LABOR OR LAND

So far we have spoken for purposes of analysis as if labor were a single homogeneous factor of production sold in a single market, and as if the same were true for land. In this event, every labor hour sold would be like every other, and every land hour of service sold like every other. Moreover, each would be bought from a single homogeneous market undifferentiated by locational or other considerations. These conditions are obviously not found in fact. The assumption that they are represents a considerable forced simplification of a complex reality, even though the general conclusions so far drawn are substantially correct. It is therefore desirable to recognize at least briefly the complications of analysis introduced by nonhomogeneity of productive factors and by the corresponding existence of disparate submarkets for each factor.

Labor as a factor of production is not a homogeneous commodity but a series of imperfect substitute commodities differentiated by skills, quality, and location. Thus there is common labor, carpentry labor, machinist's labor, various grades of supervisory and managerial labor, and so forth. Buyers do not demand labor in general but have an express demand for each type. Sellers of labor similarly offer some specific type of labor, at least at any one time. For each skill or special type of labor service there thus tends to be a separate market, with its own conditions of demand and supply, and with a quasi-independent price-determining mechanism. Similarly, the submarket for any specific skill tends to be subdivided further on the basis of location. The services of specific laborers are immediately available only in areas contiguous to their residences, and thus only to buyers whose plants are located near them. There thus tends to be for each skill a series of quasi-separate local markets, with local demand and supply conditions,[2] separated by barriers of

[2] It will be noted that the substitution conditions of labor for other factors may differ among submarkets, as well as the elasticity of the demand for the

distance and transport cost.[3] Other subdivisions of the labor market could be recognized.

So long as no additional complexity is recognized and we still suppose the existence of purely competitive buying and selling conditions in each submarket for labor, the recognition of such disparate submarkets modifies our previous conclusions in the following fashion. Within each purely competitive submarket, a provisional equilibrium wage would tend to be established, balancing the specific labor supply against the specific demand. This would result initially in a family of wage rates, for various skills and for various local markets for each skill. Thus the provisional equilibrium wage might be $2 per hour for San Francisco carpenters, $1.50 per hour for Memphis carpenters, $1.75 per hour for Memphis skilled mechanics, $2.25 per hour for Boston skilled mechanics, $300 per month for New York City junior accountants, etc. Under conditions of constant total purchasing power and adjustable wage rates, full employment could obtain in each market with such a family of potentially disparate wage rates.

This solution, however, is only provisional or transitional. Under competitive conditions there would undoubtedly be some tendency for various wage rates to adjust relative to each other, and, so far as there was mobility of supply and demand from one market to the other, to be equalized. The extent to which equalization would take place would depend basically upon the degree of mobility of the labor supply from one submarket to another. If there were perfect mobility as among skills, for example, the wage rates for all skills would tend to be equalized, presuming that every laborer sought simply the highest wage rate. Then any laborer could offer an hour of bricklaying or an hour of chemical laboratory research work with equal facility, and if bricklaying paid $3.00 per hour as compared with $2.50 for laboratory research, workers would shift to bricklay-

output which labor makes, with the result that the elasticity of demand for labor will differ among submarkets. But this will be a matter of importance, as we will see, only if there are artificial impediments to the mobility of labor among submarkets.

[3] We will not enter here into the fundamental problem of the determinants of the location of productive activity and of populations.

ing until the two rates were equalized. Similarly, the rates of pay in all occupations would tend to one mean level. The same argument may be applied to regional wage differentials, if labor were perfectly mobile and moved freely in pursuit of the highest wage regardless of where the work was offered. Then the wage rate for New Orleans steamfitters would inevitably equalize with that of steamfitters in Seattle. Supply would simply shift from market to market until the wage rates everywhere were equal.

In fact, labor is not perfectly mobile as among either occupations or areas, and thus there would tend to be, even under competitive conditions, quasi-permanent wage differentials among skills and among areas. The "natural" barriers to mobility among occupations inhere in the differences among individuals' abilities, differences in extent and type of education or training, and differences in personal preference regarding type of work. Thus not every plasterer is intrinsically equipped to qualify as a statistician (or vice versa), and few youths are financially able to acquire the education necessary to become surgeons. Further, many offspring of business managers, once given a college education, will prefer low-paid white-collar jobs to higher-paid occupations involving manual labor, because of "prestige" considerations. When barriers to labor mobility of this sort are multiplied many times, it is clear that even with every labor market competitive and with free entry to each market for all who can qualify on the basis of skill, there would tend to be sustained differentials among the wages earned in various occupations. Such mobility as personal and educational differences do not prevent would tend to limit these differentials and set them in some pattern, but even in long-run equilibrium we would have a family of different wage rates for different skills and occupations. The same may be said of regional differentials for given skills, so far as the cost of moving holds labor to given areas, or so far as traditional preference for a given state or climate induces people to forego possible advantages in pay. Thus there also tends to be a family of interrelated regional wage rates even after all natural adjustments have taken place. Regional differentials, like interskill differentials, nevertheless would tend under competitive conditions to move to some quasi-permanent

equilibrium pattern. Recognition of this fact does not seriously modify our earlier general conclusion based on the supposition of a single homogeneous labor market.

It may be worth noting that if each labor submarket were purely competitive, and if there were no arbitrary and artificial barriers to entry to any such market, wage differentials would be fixed primarily by the degree of immobility of the supply of labor, and would not reflect differences in the elasticity of demand for labor among various submarkets. Occupations where labor might exact very high wages without losing much employment would earn no more than others, unless the supply of such labor was "naturally" limited. As a consequence, the elasticity of demand for labor in specific submarkets would not be a matter of great moment.

The determination of wage differentials in practice is further complicated by the imposition of certain artificial barriers to the entry of additional labor into certain skills or areas, often because of labor-union restrictions on admission to membership, exaggerated training requirements, or licensing policies of state and municipal authorities. Where these artificial barriers to mobility exist, further causes for interskill and interregional wage differentials are introduced.

Any detailed analysis of land rents reveals the similar existence of disparate submarkets for land, separated by type, quality, or location, and giving rise to corresponding competitive rent differentials. Differentials based on the varying location and varying fertility of agricultural land have most often been emphasized, and similar emphasis might be placed upon the differentials among the rents of coal mines, oil-bearing land, agricultural land, urban site land, etc. The principal peculiarity of land-rent differentials is that they are little moderated by mobility. The locational differentiation of various pieces of land is permanent—there is no mobility from place to place, and geographic rent differentials may thus be extreme. Similarly quality differentials and differences in type, as between ore deposits and pasture land, are largely immune to mobility. A pattern of relatively permanent natural differentiation in the supply of land thus tends to fix a corresponding relatively permanent pattern of competitive rent differentials. This pattern may, of course, be

influenced by monopolistic restriction of various specific land supplies. We will not enter here into any detailed analysis of land-rent determination.

Our theory of factor pricing under universal pure competition could easily be modified to accommodate the existence of disparate submarkets for both labor and land. The same general tendencies of ultimate balance, at full employment, between factor and commodity prices and between wages and rents, remain. Thus commodity prices would still tend everywhere to equal average and marginal costs, and the ratio of wages to rents would be in balance with the marginal rates of substitution between the factors in all submarkets. Instead of a single wage rate and a single rent price, however, there would now be a family of wages and family of rents, interrelated in a complex fashion through various pairs or groups of submarkets. And the balance described would tend to be struck in each submarket as well as for the economy as a whole. Detailed analysis of such a complex balance is best pursued through the use of mathematical equations permitting the use of numerous variables and dimensions.

THE EFFECT OF COMMODITY MONOPOLIES ON WAGES AND RENTS

An outstanding aspect of the organization of the modern economy is that the selling markets for most commodities are monopolistic or quasi-monopolistic. Oligopolistic structure is most common, but there are also single-firm monopolies and markets in monopolistic competition, as well as a relatively few in pure competition. In all except the purely competitive markets, there is some tendency toward monopolistic price and output policies—toward the setting of output so that marginal cost is less than price and toward some excess profits. This has been discussed in Chapters 5 to 8. It means in effect that most firms which buy labor and land are in turn selling their outputs, of which wages and rents are the cost, under quasi-monopolistic conditions. Recognition of this fact requires the modification of our analysis of factor pricing, as begun in Chapter 10, which initially assumed that all such selling markets were purely com-

petitive. In what wise are our previous conclusions regarding rent and wages modified if we recognize that the purchasers of land and labor typically sell their outputs under monopolistic conditions? Assuming, that is, that they still buy factors in competitive factor markets, how does monopolistic commodity pricing affect wages and rents?

The impact of commodity monopolies on wages and rents has already been taken into account on pages 168-170, where the effect of a "world of monopolies" on income distribution, employment, total output, and allocation was discussed. Although these effects may be reviewed here, it is important to note that they have already been counted once. We should not make the not uncommon mistake of double counting, by measuring the effects of commodity monopoly once in analyzing commodity pricing and once more in analyzing factor pricing. In effect, firms with monopolistic selling markets follow the same general principles in buying factors as firms with competitive selling markets. They purchase labor and land (and other factors) in such proportion that the marginal rate of substitution of one factor for another is equal to the inverse of their price ratio. The *relative* prices of labor and land are thus not necessarily influenced by the existence of commodity monopoly. Whatever influence is brought to bear will be principally upon the general level of real wages and rents, upon the level of total output, and upon the allocation of resources among uses.

The potential effects of commodity monopolies on resource allocation need not be reviewed, but a brief review of over-all price and output effects may be in order. A world of commodity monopolies will individually tend to restrict output so that marginal cost is less than price and so that an excess profit margin tends to appear. Such simultaneous downward pressure on employment will tend to depress wages and rents and reduce costs. If now the flow of money purchasing power is constant at some level, the result will tend to be a reduced general level of wages and rents relative to commodity prices and an increased (excess) profit share of income going to enterprise. Total employment will be reduced so far as a smaller supply of resources is offered for employment at lower real wages and rents, but there will still be "full employment" in the sense that all wishing to work

at going rewards are employed. Commodity monopoly thus tends to reduce real wages and rents at the expense of increased excess profits, and also to affect total output by reducing the absolute level of employment which constitutes full employment. This is essentially because a monopolistic price policy in the selling markets for a commodity results in lower demand prices by firms for hired factors of production. When such a condition is general throughout the economy, the aggregate real demand for hired factors is lessened, and their prices, relative to commodity prices, are forced down. Where the flow of money purchasing power is not self-sustaining and constant, of course, there may also be an increase in involuntary unemployment.

This is as much as need be said of the effect of commodity monopolies on income distribution, so long as the firms buying factors purchase them in purely competitive factor markets, where, for example, there are many buyers of labor and many sellers of labor. The virtual impact of such commodity monopoly, however, is seldom observed in such isolation, but rather in conjunction with monopsony in the buying of factors and monopoly in the selling of them. We therefore turn to the effects of noncompetitive structure in the labor and other factor markets.

THE STRUCTURE OF LABOR MARKETS

The most significant modifications of a theory of wages and rent based on the assumption of universal pure competition must come through the recognition of the fewness of buyers of labor and land and the fewness of sellers of these factors. Monopsony or concentrated buying is extremely common in the factor markets, and especially for labor. The fewness of buyers of labor stems generally from the high concentration of the output and employment of most industries in the hands of few firms (from oligopolistic market structures) and from the fact that the labor and land markets are broken up into many submarkets, in any one of which only a small fraction of all the buyers of labor or land are situated. Because of industrial concentration, the total number of buyers of labor is not too great, and the segmentation of the labor market into many quasi-isolated parts means that only a very small proportion of all buyers purchases labor in

any one submarket. The same generalization applies to submarkets for land. This degree of concentration, which might mean that either very few or quite a few firms would purchase a given sort of labor in a given area, is the more or less "natural" result of industrial concentration and of the nonhomogeneity and imperfect mobility of the total labor supply. Such buyer concentration may be further increased or enhanced artificially by the organization of employers for the purposes of joint or concerted bargaining in the determination of wages. When this occurs, an "employers' council" or other such group may constitute an effective single-firm monopsony for a given skill within a given area. Local monopsony or oligopsony in the purchase of labor or land is effective and can remain because of barriers to the entry of competing buyers and because of the imperfect mobility of the monopsonized labor or land to other occupations or other areas.

The counterpart of concentrated buying in the various submarkets for factors is concentrated selling in such submarkets. In the case of land, this may easily result from the concentrated ownership of a given sort of land or natural resource in a given general area, resulting in monopoly or oligopoly in a particular submarket. The selling of labor has become concentrated largely via unionization on either craft or industry bases, and by the emergence of the union leadership as the single seller of the services of a large membership. Unionization has often proceeded to the point where there is a single monopolistic seller of a given type of labor for a single local submarket. In some cases, the union becomes a single seller for a group of local submarkets, when it bargains on an industry-wide basis. Unionized monopoly or oligopoly in the selling of labor may be of at least two general types: where the union is a bargaining agent for a membership to which there is essentially free entry by any laborers who wish to join, and where entry to the union is impeded by certain barriers such as arbitrary limitation, admission fees, unusual apprenticeship requirements, and so forth. In the second situation, the significance of labor monopoly is greatly increased as the mobility of labor among occupations and areas is reduced.

Under modern conditions it is not usual to find monopsonistic buying of labor with competitive selling, or monopolistic selling with competitive buying. Monopsony and monopoly, or concentrated buying and selling, are more typically combined in a bilateral monopoly situation, where a union seller faces a single buyer or organized or collusive buying group. This is the usual market setting for collective bargaining, and the one to which principal attention should be given. The general character of modern labor markets may nevertheless be explored best by considering successively the following possible cases:

1. Monopsony or oligopsony with competitive selling. This is perhaps the typical situation in the absence of unionization.
2. Monopolistic selling with competitive buying. This is the situation where a union sells to many unorganized employers.
3. Bilateral monopoly or bilateral oligopoly labor markets, where one or a few buyers deal with one or a few sellers. This is the most usual situation in the principal labor markets.

We will consider first the case of monopsonistic buying from competitive sellers.

MONOPSONISTIC BUYING OF LABOR OR OTHER FACTORS

The general effects of monopsony on the determination of the price of a good or service supplied by many competitive sellers has already been discussed in Chapter 7, and much of what was said there may be applied without serious modification to the monopsonistic purchase of a competitive supply of labor or land.

Assuming a given constant flow of money purchasing power, we have seen how with universal pure competition, including competitive buying and selling of all factors, wages and rents in each submarket for labor and land would move to an equilibrium level, in long-run balance with each other and with commodity prices at full employment. Such wages and rents may be referred to as competitive-equilibrium factor prices, and may serve as a measuring rod for assessing wage and rent behavior

under other market conditions. Given this tentative standard or norm, what can be said of the effects of monopsonistic or concentrated buying on the level of wages and rents, generally and in specific submarkets?

In examining factor-market monopsony, it is important to recognize that in the American economy there is no general monopsony or significant degree of concentration of buying for labor as a whole or for land as a whole. Monopsony and concentration of buying are peculiar to the various smaller submarkets for labor, for particular skills or occupations in particular localities; similarly it is limited to the separate submarkets for land. It follows that monopsony power is made possible (1) by the impediments to the mobility of hired factors as among various submarkets, and (2) by the impediments to the entry of added buyers of factors to specific submarkets, which keep the number of buyers at one or a few. These impediments to the entry of many buyers to a specific submarket set the stage for the exercise of monopsony power and make its exercise possible without inducing competitive buying; the imperfect mobility of factors away from monopsonized markets gives the monopsonist something to exploit and completes the picture. For purposes of our argument here we will accept the existing limitations on the entry of additional buyers to specific factor markets as generally given, and examine how an established monopsony may exploit factor immobility.

The impact of monopsony in a single submarket (or a few of them) for a factor, in an economy where there is otherwise competitive buying of factors, is most easily assessed. Let us suppose a single monopsonist in an isolated regional market for a given sort of labor—let us say for mineworkers generally in a certain coal town. The labor is supplied competitively, without union organization. Now the buyer's monopsony power presumably stems from the fact that there is a positively sloped supply curve for labor in his market, as shown in Figure 69. This means that the wage of all labor must be bid up as the amount of employment in this market is increased, according to the schedule ss'. It reflects the fact that at low wages there are some laborers, whose immobility is great, who can be employed. To secure successively more laborers—drawing them from other

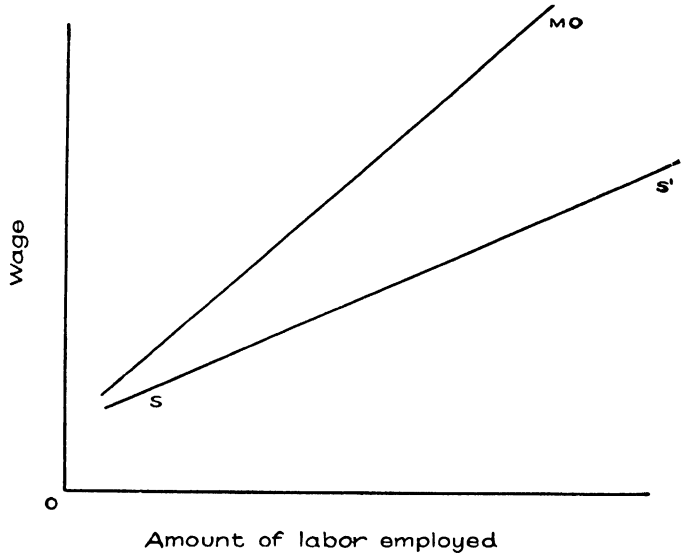

Figure 69

submarkets or keeping them from moving away—the wage rate for all labor must be bid up progressively. Thus the wage rate responds directly or positively to changes in the amount of labor purchased. Such a positively sloped supply curve would presumably be typical for monopsony purchase of labor from competitive sellers (unless there were perfect mobility, in which case supply would be perfectly elastic and there would be no monopsony power).

The impact of such monopsony power on the level of wages and employment in this submarket may be traced by reference to our analysis in Chapter 7. In setting his purchase of labor and the rate of his output, the monopsonist will refer not only to the supply curve for labor but to the marginal outlay (*MO*) for additions to labor employed, as it responds to a rising wage rate. In effect, the marginal cost of production for a firm buying labor under monopsonistic conditions rises more steeply than that of a firm buying labor competitively, since it reflects a rising wage rate as well as varying factor proportionality. And this marginal cost will lie higher at any specific wage level than that

of a competitive buyer would. As a consequence, in setting an output for which marginal cost equals marginal receipts from sales, the monopsonist will produce a smaller output, pay a lower wage, and hire less labor than a competitive buyer would. This is fundamentally because his potential influence over wages gives him a higher marginal cost at each wage rate, and thus brings him into profit-maximizing equilibrium to any given demand curve at a lower output.

His employment of and payment to labor is thus affected by the general impact of monopsony on output and price. It may also be influenced by substitution effects of monopsony. If only labor is monopsonized and other hired factors (*e.g.*, land) are not, the monopsonist will no longer hire labor and land in such proportions that the marginal rate of substitution of labor for land equals the ratio of land price to labor price. Rather he will minimize his costs where this marginal rate of substitution equals the ratio of the price of land to marginal outlay on labor. This will, in effect, restrict the employment of labor by causing it to be used in smaller proportion with land, and augment the restrictive effect already referred to. If both factors are monopsonized, of course, this restrictive effect may be largely canceled out.[4]

The total effect of monopsony on output and on factor proportions in a single factor submarket thus results in depressing wages in that market and in reducing the amount of the factor employed there, both below the competitive equilibrium level. This depression of wages tends to be accompanied by the emergence of monopsonistic excess profit. Wages in the monopsonized submarket will tend to be below those in similar markets with competitive buying, and the monopsony pressure will presumably induce some labor to move to other markets in search of better wages. Over-all employment will be reduced only if immobile factors go into idleness under monopsonistic restriction rather than shifting to other markets. The extent to which wages in the monopsonized market are depressed will increase as the

[4] Precisely, the marginal rate of substitution of A for B would be set equal to the ratio of the marginal outlay on B to the marginal outlay on A, and the net effect on factor proportions would depend on the relative elasticities of supply of the two factors.

labor supply there is less mobile and the supply curve for labor correspondingly less elastic.

The same general analysis may be applied to monopsony in land and to its effect on rents. The main difference in this case is that local land supplies tend to be entirely immobile, so that the supply curve tends toward zero elasticity. Under true monopsony of competitively sold land, land rents would thus tend to be driven toward zero.

The extent to which oligopsony, or fewness of buyers, in the market for a productive factor will produce essentially monopsonistic results will depend, of course, upon the extent of collusion or concurrence of action among the several rival buyers. Perfect collusion in oligopsony could give several large employers in a market substantially monopsony pricing of the hired factor. Effective rivalry among them would tend to raise the price, but not by a determinate amount.

We have so far spoken only of isolated monopsony in one or a few labor or other factor submarkets. What of the effect of widespread monopsony, in all or a large proportion of the submarkets, for a factor? By an extension of the argument developed at the end of Chapter 7, it would appear that widespread labor monopsony would tend to depress the general level of wages and to create monopsonistic excess profits at many places in the economy, thus seriously altering the distribution of income. The general occurrence of monopsony in submarkets also would tend to reduce mobility among such markets, reducing the elasticity of supply in any one of them. Over-all involuntary unemployment would not necessarily result if an appropriate adaptation of money income to prices took place. It might well occur, however, if with the altered income distribution—higher profits and lower wages—the ratio of consumption to income was reduced and a decline of real income was necessary to reach an equality of saving and investment. (Such an effect is also potentially attributable to monopolistic pricing in commodity markets.)

In sum, monopsony tends to bring about lower wages and higher profits, and, unless the behavior of investment and income is fortuitous, to create unemployment. And, as indicated already in Chapter 7, the result of varying degrees of monop-

sony in various places in the economy is an allocation of resources in other than the competitive pattern. All of this refers, however, to monopsony in markets where there are many sellers of any factor—a situation only historically important in many American labor markets. We must now also take account of concentrated selling in factor markets.

MONOPOLISTIC SELLING OF LABOR

Monopolistic or concentrated selling of labor arises from the organization of labor in unions, wherein the union membership appoints agents to bargain collectively with employers in determining wages, hours, conditions of work, and numerous rules and conditions surrounding employment. Unionization per se does not, of course, imply effective monopoly, so far as a significant proportion of workers may remain outside unions and constitute an unorganized competitive labor supply. Union power approaches monopoly selling power when most or all of the workers in a given submarket for labor choose to join the same union (or perhaps two or more different unions), or where the union is able to secure closed-shop or union-shop contracts with employers which bring all the employees of a given firm necessarily into a single union. The extent to which unions may secure monopoly positions in given submarkets (for given occupations or for all employees in given industries) has increased in recent years, as much more of the labor supply has become unionized and as laws have favored the development of closed-shop contracts. But the character of legislation governing unions and labor contracts is shifting, and it is difficult to say how the situation will develop from year to year. It nevertheless seems relevant to discuss the general effects of monopoly or oligopoly in the selling of labor in given submarkets.

A first thing to note about labor monopoly is that it does not necessarily imply the same thing as enterprise monopoly for a commodity. This is because the enterprise has a presumably clear-cut central motive in the desire to maximize profits, whereas the labor union has no similar generally accepted motivation. The union is not a profit-making enterprise, so that it presumably does not try to "maximize profits." The question is

what does it try to "maximize," or what are its goals in setting wages or bargaining for their determination. This question has been debated in recent specialized works on wages, but it can hardly be said to have been resolved.[5] We will content ourselves here with stating some of the principal possibilities, such as (1) that the union attempts to maximize the wage rate, either per hour or per day or week;[6] (2) that it attempts to maximize the "wage bill"—the wage rate times the number of employees or hours of employment; (3) that it tries to maximize employment, either for a fixed membership or for a flexible membership with free entry to the union; (4) that it makes some compromise among the preceding motivations; (5) that it simply tries to raise wages whenever possible and to defend against wage cuts; and so forth. It seems quite possible that different unions are guided by different motives, so that any over-all generalization is impossible.

The action of unions in setting or bargaining for wages will presumably depend strongly upon the policy the union follows regarding additions to membership. If the existent membership or leadership of a union regards its own labor submarket as a sort of "closed preserve" and in general attempts to exclude the admission of additional workers, to discourage entry, or to make high wage rates a dominant objective as compared to large employment, then its motivation is likely to turn toward maximizing the wage bill of either some or all of the membership. So long as the union can secure closed-shop contracts, it may exclude membership either arbitrarily or by securing wages high enough that the demand for labor does not exceed that available from established members. In this case, labor monopoly may have a definite tendency to raise wage rates in the monopolized submarkets above the competitive level, to distort the allocation of labor among occupations, and to distort the pattern of wage differentials. Where the union follows a policy of free admission to

[5] See John T. Dunlop, "Wage Policies of Trade Unions," *Readings in the Theory of Income Distribution*, Chap. 19; also, *Wage Determination under Trade Unions* (New York, The Macmillan Company, 1944), by the same author.

[6] Or, alternatively, to get higher rather than lower wage rates up to some rather high maximum limit.

membership, on the other hand, its motivation may vary, but precisely what it will be is difficult to say.

It is important to recognize that non-price as well as price matters will be important in the labor-selling policies of unions. Thus rules concerning techniques employed, tools used, the number of workers required on a job, hour limitations, and so forth, loom large in labor contracts. And where a union as a result of past membership policies controls an oversupply of a given sort of labor, it may well emphasize work-making rules over and above wages per se. Labor monopoly may thus influence the techniques of production employed by enterprise as well as the wage rates paid, and thus the efficiency of production generally. Since this is a large and difficult subject, we will confine ourselves here to the effect of labor monopoly on the wage rate and on employment. The impact of such monopoly, however, will vary, depending upon whether labor is sold to a group of competitive buyers, or to monopsonistic buyers in a bilateral monopoly situation. We will first consider monopolistic selling of labor to a competitive market of many small buyers.

Let us initially take the case of a single monopolistic seller of labor in some submarket—for example, a union controlling all truck drivers in a given metropolitan area—which sells labor services to a large group of competitive buyers. For the services of this sort of labor there will be some competitive market demand curve, derived from the combined output and substitution adjustments of all the buyers of this sort of labor as they react to alternative wage rates. The elasticity of this demand curve for labor reflects the elasticity of the demand for the output which these buyers make (*i.e.*, trucking service) and the substitution conditions between monopolized labor and other factors in producing this output. Such a competitive demand curve is shown as dd' in Figure 70, where W is the wage rate and Q the amount of labor employed. It shows the competitive buyers unequivocally ready to take the indicated amounts of labor at the indicated wage rates. If, now, there were a purely competitive supply of labor, there would also be a competitive supply curve, ss', and the competitive wage rate and amount of employment in this submarket would be, respectively, W_c and Q_c. The monopolistic seller of labor, however, is able to set the wage rate

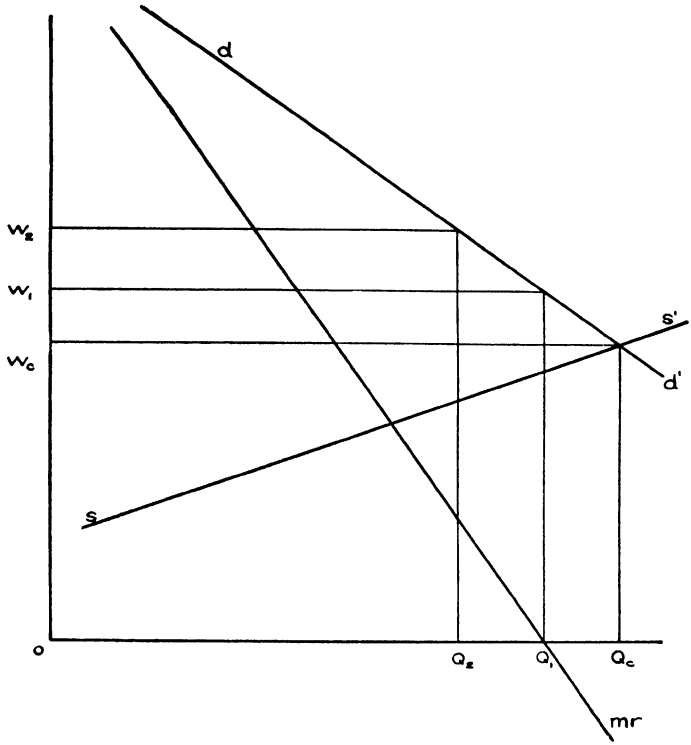

Figure 70

arbitrarily either at this level or at a higher one. What it does will depend, as suggested, upon the union membership policy and upon the motivation of the union representatives. Thus it might aim at the competitive wage rate and employment already mentioned, although at a lower rate it would presumably be unable to get a membership sufficient to meet the demand. It might also set various higher rates, depending upon the principles it adopts.

The rate W_1, at which the marginal receipts drawn to the demand curve equals zero, would maximize the wage bill of any membership, although not the wage per hour or per man. The union might thus set this rate and limit its membership to a number just sufficient to supply Q_1, or it might set this rate and sustain a larger membership through sharing of income by un-

employment benefits. Another possibility is that it might restrict membership even more severely, as at Q_2 or some lower level, and charge some higher wage, thus increasing the per hour or per man earnings up to some limit set by threat of interference or entry.

What such a monopolistic seller of labor would do cannot be predicted a priori, nor would it necessarily be the same thing in similar cases. Where the wage rate is set above the competitive level, W_c, it may be said that the seller is exercising his monopoly power in some degree. The general impact of such monopolistic pricing of labor in one or a few markets, the remainder of labor markets being competitive, is easily described. The outputs of the buyers of the monopolized labor would be restricted below the competitive level, the employment of labor in the monopolized markets reduced, and the wage rate there raised. Allocation of labor as between monopolized and competitive labor submarkets would be distorted from the ideal competitive pattern. Over-all unemployment would be increased, however, only so far as there was immobility of labor away from the monopolized submarket, or so far as aggregate purchasing power failed to adapt to the adjusted average level of wages.

Supposing that several or many union monopolies have similar motives with respect to maximizing the wage bill, or at any rate to raising the wage rate as far as consistent with maintaining some minimum of employment, it is quite possible that the wages set will differ according to the elasticity of demand for labor in particular submarkets. Where the demand for labor is very inelastic—because of an inelastic demand for the output produced, the fact that wages are a small part of cost, and the impracticality of substituting other factors for labor—the monopoly wage may be far above the competitive level. With a more elastic demand for labor, on the other hand, the discrepancy between a competitive and a feasible monopoly wage may be much less. With many labor submarkets monopolized, therefore, we may expect the varying elasticities of demand for labor to have a strong effect on differentials in wages among occupations and areas, whereas under competitive conditions these differences would not be significant. Labor monopolies tend to introduce strong arbitrary elements into wage differentials.

It is also worth noting that where a monopolistic seller of labor supplies most or all of the manpower necessary to supply an entire industry or market output, the setting of super-competitive wage rates will also result in the establishment of correspondingly super-competitive commodity prices, and that commodity output as well as labor employment will tend to be restricted. Where the demand for the final output is quite inelastic, in fact, the main burden of the monopoly wage rate may be passed directly to the consumer, with the intermediate employers suffering relatively little. In general, monopolistic labor-selling policies may have a direct impact on the relative outputs of various commodities and on the allocation of resources among uses as well as on the pattern of wage differentials as among submarkets.

Since the aggregate output of buyers in monopolized markets would be reduced, and possibly the number of buyers, their aggregate profit share as sellers of output would tend to be reduced, although the potential excess profit of any one seller of output would not necessarily be eliminated. This is because monopoly power in the selling markets of labor buyers could still operate effectively up to some high wage limit which would not necessarily be approached by the monopolistic seller of labor. Perhaps the most serious potential impact of one or a few labor monopolies would be on the structure of wage rates, on the allocation of labor as among submarkets, on the relative prices and outputs of various commodities, and in the forcing into idleness of immobile labor in the monopolized markets. It should be emphasized, however, that there is no a priori certainty that a monopolistic seller of labor will follow a monopolistic price policy. If he does, as we will see, his monopolistic gain in wage rates is almost certain to result in offsetting losses to other segments of the labor supply.

The preceding argument applies to the effect of one or several submarket monopolies in labor which exercise their monopoly power to raise wages above the competitive level. What of the potential effect of a "world of monopolies" in labor submarkets, so that all or most such markets were dominated by single monopolistic sellers of labor? Where such monopolistic sellers chose to set competitive wage rates, there would be no necessary

impact upon employment or wages. Suppose, however, that they in general set, or attempted to set, monopolistic wages above the competitive level. With differences in union motivation and differences in the elasticity of demand for labor among submarkets, a very considerable distortion of wage differentials away from the competitive pattern could, of course, be expected, with a corresponding distortion of resource allocation.

What of the effect on over-all employment? What would happen here depends upon the effect upon total money purchasing power of a concerted upward push on all wages.[7] Should total money purchasing power tend to remain constant in the face of a general upward push on wage rates to monopolistic levels, monopolistic labor prices would tend to restrict total labor employment (because of output restriction and substitution of other factors for labor), to raise the wage rate relative to the rates of reward received as interest, rents, and profits, and possibly to increase the total wage bill (if the demand for labor under these conditions is less than unit elastic). If competitive wages would just result in full employment, generally monopolistic wages would then result in unemployment and, unless unemployment benefits were very liberal, in discrimination against the disemployed labor. Justification might nevertheless be offered by interested parties that such monopolistic labor pricing was necessary to reduce the excess profit shares of monopolistic enterprises or to reduce the relative "unearned" shares of interest and rent.

It is improbable, however, that money purchasing power would remain unchanged as wages were forced up. The Keynesian argument is at the opposite extreme, to the effect that a *general* upward push on money wages would in the long run have substantially no effect on total employment or on the level of real wages, since investment would respond proportionally to all changes in wage-price level. Or, in effect, when all unions set monopolistic wages, they would, in the new equilibrium, not succeed in raising real wages or reducing employment but simply in raising the money wage-price level for the

[7] The argument is thus related to that concerning the effect on money purchasing power of a concerted downward push on wages by a world of monopsonies.

economy. This is, of course, precisely on the assumption that other factor prices move proportionally with wages and that cost-price relations remain constant. It also refers to final equilibrium tendencies rather than to phenomena of transitional periods when the economy is adjusting to recent up-pushes on wages. In such intervals, labor might, even under Keynesian assumptions, benefit from its monopolistic policies. But in the long run any single upward push on wages would fail to raise real wages, reduce employment, or cut profits or other distributive shares. The welfare of labor monopolies, as measured in terms of raising the labor share of income above the competitive level, would then lie in frequent and repeated increases in money wages and a progressive inflation of the price level.

The proportional adjustment of all prices and of money purchasing power to upward changes in the wage rate may of course be questioned, even as a long-run equilibrium tendency, and, lacking it, labor might be able to raise all real wage rates to monopolistic levels over long periods. Moreover, effects upon the structure as well as the level of wages should be taken into account in this argument. But further theoretical discussion of this issue must be left to more advanced treatments.

So far we have concentrated on monopoly affecting wages, which represent by far the largest distributive share and the dominant element in production costs for the economy as a whole. The discussion has been pursued on the implicit assumption that other factor prices are supplied under competitive conditions. In this case, competitive rents would supposedly adjust with upward or downward movements in the general level of wages, suffering from labor monopoly only so far as it was effective in raising the real wage rate. It should now also be noted that land may be sold under monopolistic conditions as well as labor, although the assumption of relatively few local land monopolies seems to accord with fact better than that of a world of land monopolies. The effect of any single land monopoly on rent in the monopolized submarket is parallel to that suggested in the case of labor; it can raise rents by restricting the use of available land. Land monopoly, however, has two distinguishing characteristics. There is a definite monopoly price (allowing a maximum total rent bill) toward which it should tend, since the

usual principles of profit maximization apply here. And the land withheld from use will generally be immobile to other submarkets, so that unemployment of available land definitely tends to result. Land or resource monopoly may be quite significant in increasing the return to passive ownership.

The preceding discussion applies explicitly to the effects of selling monopoly in factors where there is competitive buying. We must now turn to the more common case of bilateral monopoly or bilateral oligopoly in labor markets.

BILATERAL MONOPOLY IN LABOR MARKETS

With the rapid increase of unionization of labor in recent years, countered in many instances by increased organization of employers for purposes of labor bargaining, a great many labor submarkets have become highly concentrated on both the buying and the selling sides. Thus in some submarkets it is not unusual for an employers' council to bargain as a monopsonistic buyer from a union which acts as a monopolistic seller, giving rise to what is strictly a bilateral monopoly situation. In other submarkets, each of a few large firms in an industry may bargain individually (though possibly with collusion) with a single industrial union which acts as a monopolistic seller. Here we have oligopsony facing monopoly. Other labor market situations may constitute oligopoly-oligopsony, or bilateral oligopoly. At the present time, the principles governing wage determination in such markets are imperfectly understood, and abstract theory gives only very uncertain suggestions. Some consideration of the theoretical possibilities of bilateral monopoly wage determination may nevertheless shed light on this complex problem.

As in the discussion of previous models, it will be useful to distinguish between the effects of bilateral monopoly in one or a few labor markets and the effects of bilateral monopoly spread throughout the various labor submarkets. Taking first the isolated individual case, we may refer to our outline of the general possibilities of bilateral monopoly in Chapter 7. The four principal alternatives mentioned there were (1) that the monopsonistic buyer dominates and sets a low monopsony price and restricted employment; (2) that the monopolistic seller dominates

and sets a high monopoly price and (different) restricted employment; (3) that the price as set by bargaining falls indeterminately between the monopoly and monopsony limits, with indeterminate employment effects; and (4) that the buyer and seller get together and set a price for the buyer's output, allowing a maximum combined return, and then agree to divide the spoils in some fashion. In applying this analysis to labor markets, we must take account, of course, of the uncertain motivation of monopoly policy in the unionized selling of labor. That is, we can predict what the monopsonistic buyer of labor would like to do if able to set wages for his market, but we must recognize the several alternative aims mentioned on page 459 as possibly attributable to the monopolistic seller.

Where the monopsonistic buyer dominates the bilateral monopoly market, perhaps because of restricted general employment and demand for labor, we obtain the monopsonistic solution already discussed.[8] Where the monopolistic seller dominates the bargain, we get one of the several possible monopoly solutions already discussed.[9] These cases do not require further discussion. Attention is therefore appropriately focused on the case of balanced bargaining strength and on that of buyer-seller collusion.

A common situation should be that where the buyer and seller have fairly balanced bargaining strength, and where the wage is determined somewhere above the lower monopsony limit and somewhere below the upper "monopoly" limit, wherever the latter may be. In this event abstract theory can tell us little about the wage rate in a specific submarket, or how it will relate to competitive wage rates set in other submarkets. It may be the same or higher or lower. Even if it should be the same wage rate, moreover—that is, a competitive one—it is not certain that the rate of employment of labor and the output of the employing industry will also be the same as if the labor market were

[8] That is, the buyer sets a wage; at this wage the seller in essence faces a perfectly elastic demand curve for the services it offers, and the result will be as if there were many small sellers.

[9] That is, the seller sets a wage; at this wage the buyer faces a perfectly elastic supply curve for labor services; and the result is as if there were many small buyers.

competitive. In order to force the wage rate above the monopsony level, the selling union may find it necessary to restrict membership and to limit the supply of labor. And in order to force the wage rate below the monopoly level, the buying employer may tend to restrict output and employment somewhat. In effect, he may view his long-run marginal costs as rising more steeply because of the ability of the union to exact higher wages as he extends his output and employment. It follows that under bilateral monopoly the "compromise" wage may well fall near the competitive level, but the rate of employment may at the same time fall significantly below the competitive level. This, in turn, tends to restrict the output of the buying firms in industry, and probably also to allow excess profits to the buyers.

The writing of long-term wage contracts may of course dampen this tendency toward restriction of output at compromise wage levels. If the buyer views the supply of labor as perfectly elastic at the contracted wage, and the sellers similarly view it as fixed regardless of the supply of services they offer, then a competitive employment as well as a competitive wage may be approximated. But the continued anticipation of contract renegotiation may well serve as a restrictive pressure on both parties to the bargain. So far as it does, bilateral monopoly in one or a few labor markets would tend to result in some indeterminate effect upon wages but in distortion of resource allocation, restricting the use of labor in the affected industries and forcing it either to idleness or to employment in other industries.

Suppose that, instead of arriving at a compromise wage bargain, the buyer and seller of labor enter into collusion to exploit the ultimate market "to the greatest combined advantage." This is not a simple idea in the case of labor markets, since the selling union has no "profit" which may be aggregated with that of the buying firm in seeking a maximum aggregate. In fact, the most that can be said is that the union and the employer might agree jointly upon an employment of labor, a wage, an output, and a price of output which made wages attractively high and profits attractively large. Restriction of entry to union membership or employment would be an indispensable part of such an arrangement. Where it was made, restricted employment, greater-than-

competitive wages, restricted output, and high price and profits would tend to result in the affected industries.

At the present time, a large number of our most important labor submarkets are subject to bilateral monopoly bargaining conditions. What is the impact of a "world" of such labor markets upon aggregate employment, the wage level, and the structure of wage rates?

Because of the dominant uncertainty of the outcome in any one such market, the effect upon the economy of a world of such markets is very hard to predict. It would appear generally, however, that so far as unions maintain strong bargaining positions and escape the duress of monopsonistic wages they will be inclined to do so by restricting or threatening to restrict the supply of labor at lower wage rates. Further, employers will tend to restrict output and employment as a part of bargaining tactics, counting as a cost of increased output the probable rise in the negotiated wage which unions can exact. In a world of bilateral monopoly labor markets, therefore, it is quite possible (though by no means certain) that chronic underemployment may tend to result, in the sense that fewer labor hours are employed than would be under purely competitive labor-market conditions. Where money income remains relatively stable, this result is quite probable. Total output for the economy may be reduced, wages may not necessarily be better than the competitive level (although they should exceed monopsonistic wages), and monopolistic excess profits of enterprise may not be eliminated. Such a result would stable, of course, only if the disemployed laborers or labor hours were withheld from competing in the market because of union affiliation or because of the power of established unions to control entry to the labor market.

In fact, of course, a part of the labor markets are not subject to bilateral monopoly, and wages in them are determined competitively or subject to simple monopsony. In this event, the impact of the existence of many bilateral monopoly labor markets is also to distort the allocation of labor among various submarkets and to affect the structure of wages as among such markets.

One indication of the preceding analysis is that with bilateral monopoly general in labor markets, labor as a whole may have

difficulty in bettering itself so long as there is a stable money income. That is, bargaining in these markets may tend to result in wages somewhere in the competitive range but in systematic restriction of employment. If there is for other reasons a strong tendency for total money income to rise, however, labor may during such expansion periods gain close to full employment and frequent money wage increases, the restrictionist tendencies of the bargaining system being temporarily obliterated by the rising money demand for goods and services.[10]

THE GENERAL LEVEL AND THE STRUCTURE OF WAGES

The preceding discussion concerns predictable tendencies of wage and employment determination under various labor-market structures. Emphasis has been placed upon monopolistic, monopsonistic, and bilateral monopoly market situations. Evidently a very detailed empirical study of actual labor markets would be required to inform us at all fully about actual behavior in such markets. Drawing simply upon the sketchy and often uncertain predictions of a priori theory, however, it may be useful to inquire what over-all wage-employment tendencies are probably inherent in actual American labor markets.

So far as over-all employment tendencies are concerned, labor market structures are not the only determining influence. The rate of real investment activity and the relation of saving to real income are of leading importance, and the nature of the labor bargain will not *necessarily* influence these a great deal. In an economy with universally competitive factor markets we have seen that there could be alternatively tendencies toward full employment or chronic unemployment, tendencies toward persistently rising or falling money and real income. These would result from a certain behavior of investment demand and a certain relation of saving to income. Labor market structures and the character of wage determination may be said to influence the level of employment by influencing investment and the saving-income relation. The first question is what virtual influ-

[10] For a further discussion, see Fellner, "Prices and Wages under Bilateral Monopoly," *op. cit.*

ence may noncompetitive wage determination have on any preexisting tendency regarding the level of employment.

Certain potential general indications regarding over-all employment tendencies may be reviewed. First, general monopsonistic pressure on wages tends to reduce the labor share of income, but it may reduce employment only so far as more labor prefers idleness at a lower real wage. That is, the relation of real investment to real saving will not *necessarily* be modified by lower real wages in such a way as to increase involuntary unemployment. If the investment-saving relation is considerably altered, of course, involuntary employment may be increased.[11]

Second, a general monopolistic pressure on wages will tend to get wages at least up to a competitive level. But it may not tend to reduce employment except so far as monopolistic sellers of labor arbitrarily restrict the supply of labor in order to exploit an inelastic demand for labor generally. Otherwise the upward push on wages may tend to be matched by rising money income and prices, and the real employment equilibrium may be substantially unaffected.

Third, bilateral monopoly labor markets throughout the economy, although not necessarily producing systematic deviations from competitive wage levels, may result in a restriction of employment. In this case the monopsonistic buyer's tendency to restrict output and employment remains but is not necessarily offset by a yielding in real wage rates. And the monopolistic seller's virtual restriction of labor supply does not necessarily result in rising money wages and income to offset this restriction. The bilateral monopoly control may restrict both supply and demand and reduce investment, consumption, money income, and employment below competitive levels. This tentative theoretical indication requires extended further examination.

The effects of existing labor market structures on the general share of all income received by labor may be inferred in some degree from the preceding. There are some labor markets which are relatively competitive, some dominated by monopsony, some primarily monopolistic, and many with strong bilat-

[11] The ratio of saving to income may increase as personal income distribution becomes more unequal with lower real wages.

eral monopoly elements. As an average result, labor as a whole today probably protects itself against monopsonistic exploitation —against a total share of income less than the competitive level. It does this, however, by artificially restricting supply in at least some markets where it has monopoly power, and earning super-competitive shares there to offset sub-competitive shares elsewhere. Any more precise quantitative appraisal would require detailed statistical analysis. It should be noted, moreover, that a "competitive" share for labor in this sense implies mainly an avoidance of monopsonistic exploitation. There is no general reason to believe that labor via bargaining does or could eliminate those monopolistic excess profits which result from the selling monopolies of employer firms.

With a variety of labor market structures, there is, of course, a very considerable distortion of the pattern of wage rates as among various submarkets away from what might be regarded as a competitive wage-difference pattern. Monopsonistic exploitation makes wages abnormally low in some submarkets, and effective monopoly and blockade of labor entry raise wages above competitive levels by varying amounts in other cases. Significant systematic deviations of the wage-differential pattern from the competitive pattern, and from one of equal pay for equal work, tend to result from our present labor market organization.

WAGE DETERMINATION WITH FLUCTUATING NATIONAL INCOME

The preceding analysis of wages and rents is concerned primarily with the "equilibrium" tendency of these distributive shares in a situation where employment becomes relatively stable—where money income maintains a constant flow, or where, alternatively, the real income flow is constant in spite of money income and money price changes. Such equilibrium conditions are seldom realized fully, as both money and real income are continually on the move in a series of persistent fluctuations. Equilibrium wages thus represent only hypothetical goals toward which real wages tend but which they seldom attain.

Because of the tendency to persistent fluctuation, it may be instructive to consider the impact of noncompetitive wage determination upon the course of money-income and money-price

fluctuations. Two sorts of money-income movement may deserve particular attention: (1) the decline of money income which tends to occur when equilibrium employment is less than full and falling wages and prices tend to result; and (2) the rise of money income which tends to occur when investment exceeds saving even after full employment has been reached. These are crucial phases in the dynamic fluctuation of income.

Under the assumption of purely competitive factor markets, dynamic instability may result from either of these situations. Suppose that the investment-saving relation is currently such that with given prices and wages there would be equilibrium employment with 20 percent of laborers involuntarily unemployed. Then, with freely moving competitive wage rates, there would tend to be an unlimited fall of money wage rates and of all other money prices and money income, as the unemployed resources sought employment. But no price-wage decline would necessarily eliminate the tendency to unemployment so long as real investment demand and the real saving-income relation were unaffected. Or suppose that there is full employment and investment still exceeds saving by a substantial amount. Then, with competitive factor markets, money income, wages, and prices will rise indefinitely, or at least until psychological aversion to abnormally high money prices reduces real investment demand.[12] Let us see how noncompetitive wage determination may influence these tendencies to dynamic instability.

Where there is a tendency to chronic unemployment, and thus to progressively falling wages and other factor prices and falling money income, any institutional arrangement which leads to rigid money wages will permit the attainment of stable money income at an underemployment equilibrium. Thus bilateral monopoly wage contracts, or any money wage contracts which stabilize the money wage level or retard its fall will favor the attainment of money income and price stability where an underemployment equilibrium is inherent in the real investment-saving relationship. Similarly, monopolistic wage policies which resist money wage cuts as a matter of principle may promote underemployment stability. In these cases, union wage policies

[12] Or until existing real investment demand dwindles because of satiation.

may be accused not necessarily of fostering unemployment but of diverting the system from a futile pursuit of full employment via progressive price deflation. It is true, of course, that if wages were rigid and other factor prices, including rents and the quasi rents of existing capital goods, fell freely, there would be a certain substitution of other factors for rigid-priced labor which would increase unemployment. On the other hand, any arrangement which fosters progressive wage declines with falling income will tend to accelerate any existing process of money income contraction. This may well be true of monopsonistic wage policies where they are not countered by labor organization.

Where investment tends to exceed saving even after full employment is reached or approached, on the other hand, the impact of monopolistic or bilateral-monopoly wage determination may not be of such a character as to foster stability. When full employment is approximated, the bargaining position of organized labor tends to become dominant, and with income and prices at any rate rising, labor may be able to exact successive money wage increases. Each increase in turn leads in a process to higher commodity prices (including capital-goods prices) as cost functions are shifted upward. And with the general rise of prices, money investment tends to rise and the potential money income equilibrium moves higher and higher. Only a decline in real investment demand will halt the process of inflation, and this decline may be slow in developing, except as fear of lower prices later on leads to contrary speculation. Some writers have suggested that any attempt by the government to assure full employment by always providing sufficient investment will lead inevitably to progressive price inflation as unions exact progressively higher money wages in their contracts.

From the preceding discussion, it will be clear that the share of labor in the national income, although potentially observing certain equilibrium tendencies in various stable income situations, is in actuality likely to be dynamically variable with changing income. The pursuit by the wage rate of a position roughly in balance with an ever-changing income is likely to obscure any more delicate movements toward equilibrium. Analyses of dynamic processes, as well as intensive analysis of particular markets, should carry on the general suggestions developed above.

SUPPLEMENTARY READINGS

J. R. Hicks, *The Theory of Wages*, London: Macmillan & Company, Ltd., 1932.

John T. Dunlop, *Wage Determination under Trade Unions*, New York: The Macmillan Company, 1944.

Alfred Marshall, *Principles of Economics* (8th ed.), Book V, Chaps. 6-11; Book VI, Chap. 9.

Kenneth E. Boulding, *Economic Analysis*, Chaps. 9, 10.

Arthur M. Ross, "The Trade Union as a Wage-Fixing Institution," *American Economic Review*, XXXVII, Sept. 1947, pp. 566-588.

CHAPTER 14

THE PROFITS OF ENTERPRISE

Income is distributed in capitalism as it flows through a system of producing enterprises and is paid out by these enterprises as they acquire productive services from labor, the owners of land, and the suppliers of investable funds. Thus the purchasing power which enters firms as sales receipts finds its way around the circuit to emerge as purchasing power again. The bulk of the income flow at any time is evidently distributed as wages, rent, and interest. Most of those income payments are made to *hired* factors—to individuals not identified with the enterprise ownership, or to employed labor, rented land, and borrowed capital. But a share of wages, rent, and interest is ordinarily received by the enterprise ownership for the services which it supplies itself. A fair portion of the funds invested in enterprise is ordinarily owner-supplied and stands to earn an interest return for the owners, and the owners may also supply managerial or other labor as well as land.

For the share of wages, rent, and interest receivable by enterprise ownership, no definite contract may be written. The ownership may receive its imputed labor, land, and capital earnings simply as an undifferentiated aggregate left over from sales receipts after making payments to hired factors. Such an aggregate difference between the sales receipts and contractual payments of the firm is "accounting profit"; it is this difference which accountants compute periodically and designate as the "net profit" or

"net income" of the enterprise. Yet so far as this "accounting profit" is made up of wage, interest, or rent payments imputable to services supplied by the enterprise ownership (taking these services at their market value), there is logically no additional share of income discovered, but only a distinction between contractual wages, rent, and interest and imputed wages, rent, and interest.

A question therefore arises whether enterprise ownership is generally able to claim any share of income in addition to the market value of the labor, land, and capital it supplies. The enterprise owners in general take a residual share, or accounting profit, representing the difference between sales receipts and amounts paid to hired factors.[1] Does this accounting residual include more than imputed wages, interest, and rent? Any such additional share would generally be referred to as an economic profit, true profit, or *pure profit*. In this chapter we will draw together various suggestions from our preceding discussions and consider the probable existence and the sources of a true profit share of income.

THE ABSENCE OF PROFITS IN PURELY COMPETITIVE EQUILIBRIUM

The discussion of price equilibrium tendencies in purely competitive industries or economies (Chaps. 4 and 10) revealed that in those instances pure profits would move toward zero. With free entry to all industries and market-controlled prices not subject to direct influence by any seller, prices and outputs would move in the long run to a point where all sellers produced with selling price equal to the minimum average costs of production. These average costs would be made up solely of contractual wages, interest, and rents plus similar shares imputed to ownership at competitive market prices. When these costs were paid, the entire sales receipts of enterprise would be distributed and no true profit share would remain. The accounting profit or residual received by enterprise owners would be composed only

[1] After setting aside as a depreciation allowance an amount sufficient to maintain wasting capital goods.

of competitive interest on owners' invested funds and competitive wages and rent imputed to owner-supplied labor and land. Such an accounting residual has been referred to as a "normal profit" or an earning including "no excess profit"; it corresponds to having true or pure profits equal to zero. Such zero profits would characterize long-run purely competitive equilibrium, although short-run adjustments toward this equilibrium might permit of positive or negative profits for transitional periods.

It will thus appear that if pure profits do arise, this will be because of departure either from pure competition or from long-run equilibrium. Profits must in general result from monopolistic or monopsonistic tendencies in pricing, from a process of dynamic change in which long-run equilibrium is not persistently attained, or from both.

Before turning to these sources of profit, however, we should refer again to the so-called rewards of risk bearing. It was pointed out in Chapter 12 (pp. 425-429) that where any risk is involved the explicit yield on investable funds is calculated to include a premium for the assumption of this risk. Where the basic interest rate is 2 percent, capital goods may be substituted for other factors only down to a point where their yield over cost should be 6 percent, on the theory that there is a 4-percent chance of losing the investment entirely. Any successful enterprise may thus earn, in addition to basic wages, interest, and rent, an extra amount representing a premium for risk of loss successfully avoided. Part of this may be paid to creditors in addition to interest, so far as creditor securities have had to bear a premium rate. And part will ordinarily accumulate in the residual going to owners as a return on their investment.

This is an apparent extra profit in the successful enterprises. For the system of enterprises as a whole, however, such risk-bearing rewards should total zero, with the actual losses incurred tending to counterbalance the gains. If such a counterbalancing occurs, no additional share of aggregate income results, since the losers lose command over a volume of purchasing power equal to the amount of extra purchasing power that the winners gain. For the winning enterprises total receipts exceed total costs by a certain amount, but for the losers total costs exceed total

receipts by the same amount. The net relative income position of those not in a position to bear risk is thus unaffected, although risk does retard the rate of investment, and mistakes resulting in loss do reflect a certain unavoidable social loss in real output. The rewards of risk bearing for the system as a whole may thus be said to tend toward zero, and thus not to represent a source of pure profits for the economy, so long as systematic over- or under-estimation of risk does not occur. But the existence of risk and the adaptations to it do affect the distribution of income as among enterprise and result in true profits or losses for individual enterprises.[2]

Such risk-bearing rewards for successful enterprise could result even in purely competitive equilibrium (so far as this was consistent with the existence of uncertainty) and as well under monopolistic conditions. And "excessive" caution by all enterprise could result in an excess profit share for the economy as a whole. Where the rewards of excessive caution ended and the monopolistic profits of impeded entry began, however, would be an almost impossible question to answer except in pure abstraction. Correspondingly, it may be very difficult empirically to analyze the residual reward of any single enterprise or group of enterprises in such a way as to distinguish risk-bearing rewards from monopolistic or other excess profits.

The tendency of true profits to zero in a purely competitive economy has long constituted a sort of ideal, and also on occasion a justification of the essential fairness of income distribution in an enterprise economy regulated by competition. There is a serious question, however, whether the realization of such a tendency would be consistent with the effective functioning of capitalism. A system which is driven by profit-seeking motives, but in which no profit-seeker can succeed in making profits except by gambling on risk at even odds, may be a sociological impossibility. Certainly it is apparent from the history of capitalism that the efforts of enterprisers are likely to be turned principally toward escaping the profit-constricting bonds both of

[2] See F. H. Knight, "Profit," *Readings in the Theory of Income Distribution*, Chap. 27, for further comments on the risk problem.

competition and of equilibrium. This leads us to the primary considerations of monopoly and of dynamic change as the source of pure profits.

MONOPOLY AND MONOPSONY PROFITS

The discussions of noncompetitive pricing of goods in Chapters 5 to 8 and of noncompetitive factor pricing in Chapter 13 have suggested that monopolistic and monopsonistic market structures may permit enterprise to earn an excess or pure profit reward more or less persistently. Even in long-run equilibrium, the selling firms may be able to maintain a gap between price and full average cost and thus to divert a share of total purchasing power to themselves in return for "nothing." This could hold in either single-firm monopoly or oligopoly selling, and in monopsony or oligopsony buying. In any instance, however, the ability to earn monopoly profits depends primarily on the blockade of the entry of additional enterprises by those already established. Free entry would tend to eliminate monopoly or monopsony profits. But the existence of legal, institutional, and economic impediments to entry [3] permits the maintenance of at least some excess profits in many fields. Such profits, which may also be referred to as the earning power of whatever it is that firms possess which discourages entry, are a distinct income share and identifiable as pure profits. They will ordinarily be reflected in an accounting net profit which when expressed as a percentage of owners' investment is larger than the interest rate plus a normal risk premium. Even this will not appear, however, if the investment has been so valued as to include the capitalized present value of monopoly earnings attributable to existing blockades to entry.

It is thus an almost inevitable tendency of capitalist enterprise to seek monopoly positions continuously through time by securing buying or selling market positions protected from the ultimate in free entry. The possibilities of resource monopolization, the patent and trademark laws, the issuance of monopoly franchises by governments, and simple high concentration within

[3] Cf. pp. 148-149 above.

industries all offer avenues for gaining monopoly positions. But since blockades to entry are seldom completely effective, since the antitrust laws oppose them in some degree, and since dynamic change permits new monopolies to destroy old ones, the accumulation of monopoly may not continue indefinitely but may instead reach a peak or limit. At any current time, however, monopoly excess profits are a significant share of total income.

DYNAMIC CHANGE AND PROFITS

Monopoly is not the only source of pure profit, however. Dynamic changes in income and innovations of either technique or product may also give rise to a distinct share of income to enterprise.

The potential effect of a fluctuating aggregate money income is quite apparent. As money income expands, even from an initial profitless equilibrium of prices and costs, the money prices of both commodities and productive factors tend to rise, since their supplies in general are less than perfectly elastic. But in general commodity prices will tend to rise more rapidly than factor prices, so that a transitional margin between prices and costs is created. This is in general because of the lagging response of increases in output to increases in employment and in money payments to factors, and also because of the relative curtailment of consumer-goods output in an investment boom once full employment has been reached. As a consequence, excess profits tend more or less automatically to arise in periods of secularly or cyclically expanding money income and inflating prices. Conversely, periods of contracting money income and deflating prices tend to produce losses, or a reduction of monopolistic excess profits which would otherwise be earned. These profits and losses are essentially of a short-period variety, accountable as the result of a process of adjustment in pursuit of new money-price equilibria. It may be argued that in the long run the losses of downswings exactly counterbalance the gains of upswings, so that on the average no genuine profit share is created. This is true only if the downswings are just as great as the upswings—if there is neither progressive deflation nor progressive inflation—and if the lag of factor prices behind commodity prices is

the same in upswing and downswing phases. Otherwise, a systematic average profit or loss for enterprise may be created. And there is at any rate a tendency to periodic distortion in the pattern of income distribution.

Since individual enterprises are generally unable to control the fluctuations of aggregate money income which occur, the profits and losses arising from this source may be appropriately characterized as windfalls outside their control. This is not so, however, of profits created by dynamic changes in product or technique, since these are in general purposefully introduced by enterprise precisely in pursuit of a pure profit.

The equilibrium to which a competitive economy would tend rests on the assumption of given techniques of production, from which emerge given cost curves for all goods, and of a given regimen of goods for production. With these things given, a long-run general price equilibrium with universal pure competition would eventually be reached, at which all profits would tend to zero. With monopolistic elements present, this general price equilibrium might permit of excess profits in certain sectors of the economy, but each monopoly profit would have distinct limits. If, now, the techniques of production are changed in one or more lines, the real costs of production there are presumably reduced. Output therefore tends to expand in the affected lines and to be adjusted elsewhere, as the economy seeks a new general price equilibrium corresponding to the new technological data. When the new equilibrium is finally attained, competitive profits should again be zero, and monopolistic profits not necessarily larger than before. But during the process of movement toward a new equilibrium—and this process takes time—the initial innovators who first reduced their costs should enjoy a period during which they can earn excess profits. Until there is sufficient imitation of or entry into the use of the new technique, and until prices are thus driven down to the level of the new costs, the innovators tend to reap an extra reward as a result of their pioneering.

Under universally competitive conditions, the period during which they could thus receive extra profits might not be too long. But when patent monopolies can be obtained for a period of years, innovators may be able to create temporary monopoly

positions which exclude entry and preserve their advantages for a considerable time span. The combination of dynamic change and of blockade of entry to new techniques may thus create a distinct extra profit share. So far as enterprise will continually be engaged in developing and protecting technical improvements, this share may persist regularly through time, as the economy pursues an ever-moving equilibrium tendency. Here is the outstanding example of systematic and successful effort by enterprises to escape competitive equilibrium.

The same general argument may be applied to innovations of new products, which change the structure of demand for various goods, give rise to a new general equilibrium tendency, and create a transitional opportunity for excess profits to the innovators.

The profits of dynamic change forced by innovators are probably the principal source of a genuine and distinct profit share of income in a capitalist economy. Together with monopoly and monopsony gains they make up most of the total of such profits, the remainder representing any net windfalls resulting from income movements (some of which may be generated by the investment associated with innovation). It is also clear that profits to monopoly and to innovation are interrelated, in that innovation may open the road to new monopoly as well as tend to reduce the value of old monopoly. The pursuit of at least transitional monopoly profits may indeed be a primary incentive to innovation.

Innovations of technique and product, whether or not assisted by monopoly positions established through innovation, are a principal source of pure profit in the real economy. This admitted, the question arises whether there is not a fourth sort of productive service or "factor of production" which earns this share of income. The profits of innovations indeed go to the entrepreneur-owner group as a reward for its astuteness in changing techniques and products, and also provide a payment to that special category of labor devoted to devising possible ways of changing products and techniques. They may thus serve as a desirable incentive reward for invention and innovation, and may promote progress in the aggregate output of the economy. Because the relation of effort to reward seems highly un-

predictable and erratic in this case, however, the recognition of a "fourth factor of production" is not necessarily appropriate. It is sufficient to recognize that a share of the residual pure profit going to enterprise may result from purposive endeavor on the part of enterprise in changing techniques of production and the design of products, and that some sort of rather uncertain relation of innovation-reward to innovation-effort may thus result.

When profits are considered in this light, it appears that if they are not a payment for money, land, or labor (except that special human effort involved in innovation), they may at any rate in part be a return earned as a result of potentially constructive human effort. But innovation profit and monopoly profit are almost inextricably intertwined. And although the profits of innovation per se tend to perform a desirable function in promoting progress, their emergence in a setting permitting of well-fortified and long-perpetuated monopoly positions could retard progress and result in a total profit share larger than necessary to stimulate progress.

THE RECIPIENTS OF PROFITS FROM INCORPORATED BUSINESS

The traditional justification of economic profit in a capitalist economy has been that it furnished a necessary incentive for the enterprise system to function. At least the possibility of rewards over and above a normal interest return to funds invested has been held desirable as a means of inducing both rational allocation of resources among uses and progress in techniques and products. In equilibrium, the firm should try to maximize profits, and if the maximum is zero, it should still strive toward it. Dynamically, enterprises should be induced to make innovations of technique or to introduce new products because they promise larger profits. The potentiality of profit and the desire to maximize it should serve as a rationalizing force in enterprise action.

Considerable interest has therefore been expressed in the development of business organization to a point where the owner-investors do not make the decisions which affect enterprise profits, and where the managers who do make these decisions are not primarily owners destined to benefit directly from in-

creased profits. In very large corporations, which do over a third of all business done in America today, it is often true that the decision-making management is largely independent of the dividend-receiving shareholders, in that the shareholders, in practice, exercise little influence over the selection and retention of the management. In these cases, the shareholder becomes a passive investor, who receives part or all of the profit reward of the enterprising ability of the entrepreneur-manager; the managers, rewarded principally by wages, do not necessarily receive the pure profit rewards which should provide them with the incentive to maximize profits. Where this situation obtains, it has been suggested that controlling managements may not be primarily concerned with profit maximization as a guide either to current price-output adjustments or to policy respecting innovation.

The general facts of separation of ownership from control or entrepreneurship are fairly clear. The phenomenon is apparently significant in something like half of the largest corporations. Whether or not the principles upon which management decisions turn have been seriously affected by the separate identity of profit makers and profit receivers is still a subject for speculation and investigation. And the extent to which hired managers are paid a salary which includes some pure profit earned by reason of innovating ability can also bear investigation.[4]

THE DISTRIBUTION OF INCOME—ETHICAL ASPECTS

With the preceding discussion of profits, the general principles governing the distribution of income among productive factors have been fully outlined. Under various sorts of market organization, we are able to see the general determinants of the relative rates of pay received by labor, land, capital, and enterprise. The aggregate share received by each factor, whether under competitive or noncompetitive market conditions, will depend, of course, upon the amount of it available and employed

[4] For a discussion of these matters and also of the analysis of profits generally, see R. A. Gordon, "Enterprise, Profits, and the Modern Corporation," *Readings in the Theory of Income Distribution*, Chap. 29.

as well as upon the rate of pay per unit of service. It is generally thought that wages and salaries paid to nonowners may constitute roughly two thirds of national income, and that imputed wages of owner-managers add to this wage share to some extent. Rent may constitute from 10 to 15 percent of national income, and interest and profits the balance. These are extremely rough proportions, subject to systematic secular and cyclical movement.

Whatever the functional distribution of income under either competitive or monopolistic conditions, it has no necessary a priori ethical content. The fact that a free-enterprise economy distributes income in this way does not mean that the distribution is just or unjust. Accepting the postulates of a capitalist system—principally private property in productive wealth and the right to receive property incomes in the form of rent, interest, or profit—it is perhaps expedient that incomes should be distributed about as they are, in order that the economic system function effectively. But preference for a competitive rather than a monopolistic determination of income shares must result from the essentially arbitrary adoption of a value judgment concerning desirable interpersonal relationships with respect to income. Such value judgments are necessary in social conduct, and are freely, if seldom unanimously, adopted. The ethical evaluation of income distribution within capitalism must thus be rooted in political philosophy rather than in technical economic analysis.

This is equally true of the ethical evaluation of capitalistic as compared to socialistic income distribution. The salient aspect of capitalistic distribution is that property ownership is the basis of personal income in the form of rents, interest, and profits, whereas in full socialism the bulk of personal income should theoretically be in the form of wages for labor, other income shares being arbitrarily allocated by the state. The choice must essentially rest upon complex political value judgments. The property incomes of capitalism are the source of argument principally because the property from which they stem is rather unequally distributed (through chance, inheritance, accumulations of wealth based on prolonged historical discrepancies in personal income distribution, etc.). This unequal distribution of property in turn leads to a degree of inequality in the distribu-

tion of incomes among persons—in extremes of poverty and wealth—much greater than would result from differences in personal wage-earning ability.[5] And this in turn raises some conflicts with the precepts of political democracy. The resolution of this difficulty is not obvious, and the current complex conflict of ideologies reflects the dilemmas encountered when a real solution is sought.

Modification of personal income distribution via progressive income taxation is a principal expedient measure followed today in capitalistic countries. Inheritance taxes have the effect of checking the progressive accumulation of wealth from one generation to another. And it is felt by many that reduction or elimination of monopolistic profits and monopolistic rents would perceptibly reduce the inequality in personal income distribution. Discussion of the ethical and social problems of income distribution, however, must be left to other works.

SUPPLEMENTARY READINGS

ROBERT A. GORDON, "Enterprise, Profits, and the Modern Corporation," *Readings in the Theory of Income Distribution*, Chap. 29.

FRANK H. KNIGHT, "Profit," *ibid.*, Chap. 27.

——, *Risk, Uncertainty, and Profit*, Boston: Houghton Mifflin Company, 1921.

[5] See Mary Jean Bowman, "Personal Income Distribution in the United States," *Readings in the Theory of Income Distribution*, Chap. 4.

CHAPTER 15

CONCLUSION

This volume has presented a systematic though elementary analysis of several principal aspects of the function of a capitalist economy. We have investigated the determination of commodity prices and outputs and of the allocation of resources among alternative uses; the determination of the aggregate level of employment and output; the manner in which income is distributed among various functional groups. If the student considers the several phases of the preceding analysis in concert, he will observe a certain body of abstractly derived or a priori predictions of the working of a free-enterprise economy in the respects emphasized.

Yet it should be very strongly emphasized that these predictions and the analysis from which they arise are extremely simplified and are potentially reliable only within a substantial range of error. In the first place, such an abstract analysis essentially proceeds by setting up certain assumptions concerning human psychology and the market structures within which human action takes place, and then by deducing what should happen in the assumed situation. The assumptions are carefully drawn, and every attempt is made to have in them an accurate if simplified representation of reality. Some of these assumptions, however, may be inaccurate, or so oversimplified as to be unable in a single average to take account of a range of conditions which occur in fact. Thus the assumption that businessmen

always try to maximize profits may be a fair guess at the central tendency, but the actual motivation in specific cases may be different, or at any rate more complex. In the case of monopolistic sellers of labor, we really do not know what to assume concerning motivation and thus can only make alternative guesses. Or in distinguishing among market structures, only the broadest differences are recognized, and other individual differences of potential significance to pricing are entirely overlooked in the assumptions. With both oversimplification and chance of error affecting the assumptions upon which the analysis is based, it is clear that the predictions arrived at cannot be precise estimates of real behavior but only rough indications of a central tendency. Much more detailed abstract analysis, or alternatively detailed empirical observation, would be required to gain a more precise notion of what happens.

There is another reason why recourse to empirical investigation should be essential in many instances. When we encounter oligopolistic markets, or markets with bilateral monopoly or bilateral oligopoly, abstract analysis tells us that the outcome is indeterminate within a significant range. This applies both to commodity price-output determination and to factor pricing in labor and other markets. In these instances, we are presented not with the prediction of a central tendency but with a range of alternative central tendencies and perhaps some reason for believing one to be more probable than another. This dilemma can be resolved best by detailed empirical examination of behavior and its rationale in particular cases. A priori, it would seem that economic activity in a capitalist economy may lack a precisely observed law of behavior and be subject to a certain random tendency.

This leads us to another limitation of the particular sort of abstract analysis followed above. For purposes of argument we have assumed that the significant determining variables or variable relationships—prices, costs, demand curves, cost and supply curves—are known to the principal actors in the scene, or at any rate are estimated with sufficient confidence that they are acted upon. Thus we have supposed "given data"—given either by foreknowledge or by "reliable" estimate. In actuality, of course, the magnitudes of many strategic variables are highly uncertain;

as a result they can be estimated with no great confidence, and the reactions to these estimates may be modified by the recognized uncertainty to which they are subject. In the preceding analysis we have taken substantially no account of the idea of a range of alternative estimates of strategic variables, of the effect of such a range on decision making, or, more important, of the consequences of behavior based on erroneous expectations. This is the province of dynamic process analysis, into which we have not entered here.

A final limitation of the preceding analysis is thus that it is an analysis of equilibrium tendencies rather than of the process of change through time. Most of the preceding indicates little about the process of economic activity through time, but rather emphasizes the equilibria toward which this activity tends over a period of time. It describes the path of the rabbit which the hunter pursues rather than the path followed, step by step, by the hunter. It tells us something, but certainly less than all, of the character of economic activity. Dynamic process analysis is a logical further step in the study of economic theory.

Even in the analysis of stationary equilibria, of course, the preceding is elementary in the extreme, deliberately neglecting many advanced or detailed facets of theory, by-passing entirely alternative formulations, and emphasizing only slightly the formal mathematical aspects of theoretical solutions. It is hoped, however, that the volume will have served as a useful introduction to abstract economic analysis.

INDEX

Adjusted dollars, 373*f.*
Allocation of resources
  explanation of, 6*f.*
  in monopoly, 151*f.*, 164*f.*
  in monopsony, 229, 236*ff.*
  in oligopoly, 220
  in pure competition, 127*ff.*
  summary of, 259*f.*
Aluminum, 172*ff.*
Antitrust laws, 185, 212
Automobiles, 186*n.*, 204
Average costs, 69*ff.*

Bain, J. S., 18*n.*, 175
Bank rate of interest, 367*ff.*, 388*ff.*
Bilateral monopoly, 234*f.*, 433, 466*ff.*
Bilateral oligopoly, 234*f.*
Böhm-Bawerk, E., 365
Boulding, K., 69*n.*, 88*n.*, 94, 117*n.*, 173*n.*, 175, 285*n.*, 288*n.*, 475
Bowman, M. J., 487*n.*
Burns, A. R., 135, 187*n.*
Business saving, 371*ff.*

Capital
  definition of, 271*f.*, 312*ff.*
Capital (Cont.)
  demand for, 318*ff.*
Capitalism, 4*f.*
Cartels, 185
Cash balances, 376*ff.*
Ceteris paribus, definition of, 22
Chamberlin, E. H., 16*n.*, 50*n.*, 53*n.*, 55*n.*, 60, 135, 213*n.*, 221, 247*n.*, 248*n.*, 252
Cigarettes, 197*f.*
Circular flow, 276*ff.*
Classification of industries, 43*ff.*
Collusion, 56, 184*f.*, 187*f.*, 213*f.*, 233*f.*
Complementarity, 39
Conference on Price Research, 94
Consumer choice, 27*ff.*
Consumer finance, 360*ff.*
Corporation profits, 484*f.*
Cost
  average, 69*ff.*
  definition of, 63*ff.*
  fixed, 65*ff.*
  long-run variation in, 82*ff.*
  real, 2*n.*, 61, 444*f.*
  short-run variation in, 66*ff.*
  variable, 65*f.*, 76*ff.*

Cost curves
effect of monopsony on, 223
envelope, 90*f.*
long-run, 82*ff.*
shifts in, 90*f.*
short-run, 66*ff.*

Dean, Joel, 80*n.*
Demand curves
definition of, 12*ff.*
derivation of, 28*f.*
elasticity of, 30*ff.*
for investable funds, 328*ff.*, 353*f.*, 356, 362, 364, 379, 382*f.*
for labor, 288*f.*, 290, 293, 461
for sellers' outputs, 41*ff.*
groups of, 36*ff.*
industry, 19*ff.*
in monopolistic competition, 242*ff.*
in monopoly, 44*ff.*, 142*ff.*
in monopsony, 224*ff.*
in oligopoly, 55*ff.*, 180*f.*, 191*f.*, 196
in pure competition, 47*ff.*, 102*ff.*
kinked, 183*f.*
negative slope of, 27*ff.*
shifts in, 25*f.*, 40
Demand schedules, 12*ff.*
Dennison, H. S., 221
Destructive competition, 124
Differentiation of product, 15*f.*, 50*f.*, 131*f.*
Diminishing returns, 69*n.*
Discounts, 199
Discrimination, 173, 173*n.*
Diseconomies of large scale, 85*f.*
Divisibility of factors, 87*f.*
Divisibility of plant, 78*f.*
Drury, H. B., 221
Dunlop, J. T., 459*n.*, 475
Dynamics of monopoly price, 157*ff.*

Economies of large scale, 84*ff.*
Efficiency
in monopolistic competition, 251
in monopoly, 153*f.*
in oligopoly, 208*f.*
in pure competition, 120
summary of, 261
Elasticity of demand, 30*ff.*, 48*n.*
Ellis, H. S., 437
Employment
and commodity prices, 255*ff.*
and monopoly, 168*ff.*
and monopsony, 230
and oligopoly, 220
and wage policy, 464
conditions for full, 292, 295, 301
equilibrium in, 407*ff.*
less than full, 298, 301
"over," 439
Entrepreneurship, 273*f.*, 483*ff.*
Entry
barriers to, 139*f.*, 179
definition of free, 99*n.*
excessive, 124
threat of, 155*ff.*, 189*ff.*
Envelope curve, 90, 90*n.*
Ethics of distribution, 485*ff.*
Excess profits, 147*ff.*

Factors of production, definition of, 270*f.*, 311*f.*
Federal Reserve System, 389*ff.*
Fellner, W., 239, 348*n.*, 365, 470*n.*
Fewness of buyers, 222*ff.*
Fewness of sellers, definition of, 54, 178
Fisher, I., 365, 437

Fixed cost, definition of, 65*f*.
Fixed factor, 65
Formula pricing, 186, 193*f*.

Galbraith, J. K., 221
Gasoline, 199
Gordon, R. A., 485*n*., 487
Government debt, 434*ff*.
Gross investment, 325*ff*., 361*ff*., 403*ff*.

Haberler, G., 437
Hall, R. L., 194*n*.
Hamilton, W., 221
Hansen, A. H., 267*n*., 342*n*., 348*n*., 365, 401*n*., 403*n*., 435*n*., 437
Hart, A. G., 310, 429*n*.
Hayek, F. A., 99, 322*n*.
Hicks, J. R., 28*n*., 60, 239, 288*n*., 307*n*., 310, 475
Hitch, C. J., 135, 194*n*., 252
Hoarding, 278, 300, 376*ff*.
Homogeneity of product, 15

Imputed value, 63*f*.
Independence of sellers, criteria of, 182*n*.
Industrial revolution, 130*f*., 344*f*.
Industry
  classification of, 43*ff*., 177
  definition of an, 14*ff*., 59*n*.
Inferior goods, 40*n*.
Innovation, 483*f*.
Interest cost, 325*f*.
Interest rate, 317, 326*ff*., 367*ff*., 384*ff*., 420*ff*.
Interest share in income, 420*ff*.
Investable funds
  and liquidity preference, 376*ff*.
  bank supply of, 367*ff*., 388*ff*.
  demand for, 328*ff*., 353*ff*.
Investable funds (Cont.)
  necessity for, 272*f*., 313
  requirements of, for consumer finance, 358*ff*.
  supply of, 366*ff*.
Investment
  gross, 352*ff*., 361*f*., 403*ff*.
  net, 314, 329*f*., 335*ff*., 348*ff*., 361*f*., 403*ff*.
  reinvestment, 329*ff*., 354*ff*., 362, 401*ff*.
  total, 325*ff*.
Isoquants, 286*f*., 319*f*.

Kahn, R. F., 55*n*., 175
Keim, W. G., 112*n*., 249*n*.
Keynes, J. M., 310, 329*n*., 365, 378*n*., 412, 437
Kinked demand curve, 183*f*.
Knight, F. H., 437, 479*n*., 487

Labor
  conditions of supply of, 442*f*.
  definition of, 271
  structure of markets for, 451*ff*.
  submarkets of, 445*ff*.
Laissez faire, 99, 132
Land
  conditions of supply of, 443*f*.
  definition of, 271
  monopoly in, 365*f*.
  submarkets of, 445*ff*.
Lange, O., 310
Large-scale production, 84*ff*.
Leontief, W., 214*n*.
Lerner, A. P., 129*n*., 135, 165*n*.
Lewis, B. W., 268
Liquidity preference, 376*ff*.
Long period
  cost behavior in, 82*ff*.
  definition of, 63, 65*f*., 82, 96

Lundberg, E., 437
Lutz, F. A., 437

Machlup, F., 43*n*.
Marginal cost, 88*ff*., 127*ff*., 151*f*., 236*ff*.
Marginal outlay, 225*ff*., 423
Marginal rate of substitution, 284*ff*.
Marginal receipts, 102*f*., 143*f*.
Market classification, 43*ff*.
Marshall, A., 9, 60, 94, 475
Mason, E. S., 44*n*.
Meade, J. E., 135, 252
Money, character of, 376*ff*.
Monopolistic competition
  character of, 240*ff*.
  definition of, 50*ff*., 50*n*.
  examples of, 241, 249*f*.
  pricing in, 242*ff*.
Monopoly
  and capital goods, 429*ff*.
  and employment, 168*ff*., 256*ff*.
  bilateral, 234*f*.
  definition of, 44*ff*., 45*n*., 138*f*.
  demand in, 44*ff*.
  dynamics of, 157*ff*.
  effect of, on wages and rents, 449*ff*.
  in selling of labor, 458*ff*.
  output in, 150*ff*.
  pricing in, 140*ff*.
  profits in, 147*ff*., 168, 480*f*.
  selling costs in, 158*ff*.
Monopsony
  and capital goods, 429*ff*.
  basis of, 228*ff*.
  demand in, 224*f*., 230*ff*.
  effect of, on wages, 453*ff*.
  price in, 226*ff*.
  profits in, 226*f*., 480*f*.
Mutatis mutandis, definition of, 22
National Bureau of Economic Research, 80*n*.
National Resources Committee, 54*n*.
Natural monopoly, 85, 174*f*.
Nelson, S., 112*n*., 249*n*.
Net investment, 314, 329*f*., 335*ff*., 348*ff*., 361*f*., 403*ff*.
Nicholls, W. H., 234*n*., 235*n*., 239
Nonprice competition, 200*ff*., 247*ff*.
Normal profit, 64
Normative behavior, 120*ff*., 125*ff*.
Norris, R. T., 60
Nourse, E. G., 221

Oligopoly
  advertising in, 205
  allocation in, 220
  bilateral, 234*f*.
  cartels in, 185
  collusion by, 184*f*., 187*f*., 213*f*.
  definition of, 53*ff*.
  demand curves of, 180*f*., 191*f*., 196
  effect of entry in, 189*ff*., 215*f*.
  examples of, 54, 180
  independence of sellers in, 182, 182*n*., 212
  indeterminacy of, 55*ff*., 182
  interdependence of, 181*f*., 209*ff*., 213*f*.
  market shares in, 197*ff*.
  nonprice competition in, 200*ff*.
  price leadership in, 50, 185*f*., 189
  price policy in, 180*ff*., 187*ff*., 212*ff*.
  price results in, 207*ff*., 217*ff*.
  profits in, 219*ff*.
  reasons for, 178*f*.
  rivalry in, 181
  subcategories of, 177*f*.

Oligopsony, 233*f.*
Opportunity price, 63*f.*
Optimum output, 72*f.*, 129*f.*

Patents, 140
Patinkin, Don, 214*n.*
Pecuniary economies, 88
Population growth, 335*ff.*
Price analysis, function of, 7*f.*
Price jurisdiction, 137
Price leadership, 56, 185*f.*, 189
Price rigidity, 121, 183, 214*f.*, 264*ff.*
Pricing
  in monopolistic competition, 242*ff.*
  long-run competitive, 113*ff.*
  monopolistic, 140*ff.*
  monopsonistic, 224*ff.*
  oligopolistic, 180*ff.*
  short-run competitive, 99*ff.*
Process analysis, 490
Profits
  accounting of, 475
  and corporate business, 484*f.*
  as distributive share, 273*f.*, 476*ff.*
  distinguishing of, from risk premium, 478*f.*
  excess, 147*ff.*
  from dynamic change, 481*ff.*
  in competitive equilibrium, 120, 129, 477*ff.*
  in monopoly, 147*ff.*, 168, 480*f.*
  in monopsony, 226*f.*, 480*f.*
  in oligopoly, 219*f.*
  normal, 64
  pure, 476
  resulting of, from fluctuating income, 302*ff.*, 481*f.*
Progressiveness
  in monopoly, 159*f.*
Progressiveness (Cont.)
  in oligopoly, 204*ff.*, 208*f.*
  in pure competition, 121*f.*
  summary of, 262*f.*
Public utilities, 174*ff.*
Pure competition
  conditions for existence of, 130*ff.*
  definition of, 47*ff.*, 99*f.*
  demand in, 47*ff.*
  factor pricing in, 280-295
  instability in, 122*ff.*, 133*f.*
  long-run pricing in, 113*ff.*
  normative properties of, 120*ff.*
  price results of, 119*f.*
  short-run pricing in, 101*ff.*
  significance of, 97*ff.*

Quasi-rents, 106, 272, 313

Real cost, 2*n.*, 61, 444*f.*
Reinvestment, 329*ff.*, 354*ff.*, 362, 401*ff.*
Rent
  considerations affecting, 441*ff.*
  definition of, 271
  differentials, 445*ff.*
  source of, 275
Reynolds, L. G., 123*n.*
Risk premiums, 425*ff.*, 478*f.*
Road to Serfdom, 99
Robbins, L., 9
Robinson, E. A. G., 94
Robinson, Joan, 119*n.*, 226*n.*, 239
Ross, A. M., 475

Sales policies, determinants of, 10*ff.*
Saving
  definition of, 371*ff.*
  relation of, to income, 371*ff.*, 399*ff.*, 404*f.*
Scale of firm, 62, 82*ff.*

Schultz, H., 29*n.*, 60
Schumpeter, J. A., 9, 167*n.*, 175, 310, 334*n.*
Selling costs
  in monopolistic competition, 247*ff.*
  in monopoly, 158*ff.*
  in oligopoly, 200*ff.*
  relation of, to output, 93*f.*
  summary of, 261*f.*
Short period
  and cost behavior, 64*ff.*
  definition of, 63, 65*f.*, 96
  pricing in, 99*ff.*
Single-firm monopoly. *See* Monopoly
Smith, Adam, 99, 111
Sombart, W., 9
Specialization, 84*ff.*
Stability
  in pure competition, 133*f.*
  of money income, 395*ff.*
  summary of, 263*ff.*
Stationary state, 329
Steel, 210*ff.*
Stigler, G. J., 18*n.*, 35*n.*, 46*n.*, 60, 90*n.*, 117*n.*, 135
Structure of production, 315
Substitution
  among factors, 284*ff.*, 319
  among goods, 13*f.*, 456*ff.*
  between capital and other factors, 320*ff.*, 423*f.*
Supply curves
  for commodities, 109*f.*, 113*ff.*
  for funds, 366*ff.*, 372*f.*, 381, 383, 400, 402
Supply curves (Cont.)
  for labor, 282, 291*f.*, 294*f.*, 299, 455
Sweezy, Paul, 184*n.*

Taylor, F. M., 310
Technological progress, 324*ff.*
Time intervals
  in income analysis, 369*ff.*
  in pricing, 95*ff.*
Triffin, R., 60, 60*n.*

Utilization of plant, 62

Variable cost
  behavior of, 76*ff.*
  constancy of, 80
  definition of, 65*f.*
Variable factor, 65
Variable proportions, 69*n.*, 77*ff.*
Variables, definition of, 21*f.*
Veblen, T., 9
Velocity of money, 377*f.*
Very short periods, 95*ff.*
Viner, J., 94, 119*n.*

Wages
  considerations affecting, 441*ff.*
  definition of, 271
  determinants of, 449*ff.*
  differentials of, 445*ff.*
  distinguishing of, from rents, 274*f.*
  real, 298*ff.*
Wallace, D. H., 172*n.*, 268
Wealth of Nations, 99
Wicksell, K., 365
Wilcox, C., 175