BUSINESS CYCLES AND ECONOMIC GROWTH

Duesenberry

ECONOMICS HANDBOOK SERIES

Business Cycles and Economic Growth

ECONOMICS HANDBOOK SERIES

SEYMOUR E. HARRIS, EDITOR

Burns · SOCIAL SECURITY AND PUBLIC POLICY

Carlson · ECONOMIC SECURITY IN THE UNITED STATES

Duesenberry · BUSINESS CYCLES AND ECONOMIC GROWTH

Hansen · A GUIDE TO KEYNES

Hansen · ECONOMIC ISSUES OF THE 1960S

Hansen · MONETARY THEORY AND FISCAL POLICY

Hansen · THE AMERICAN ECONOMY

Harris · INTERNATIONAL AND INTERREGIONAL ECONOMICS

Henderson and Quandt · MICROECONOMIC THEORY

Hoover · THE LOCATION OF ECONOMIC ACTIVITY

Johnston · STATISTICAL COST ANALYSIS

Kindleberger · ECONOMIC DEVELOPMENT

Lerner · ECONOMICS OF EMPLOYMENT

Taylor · A HISTORY OF ECONOMIC THOUGHT

Tinbergen and Bos · MATHEMATICAL MODELS OF ECONOMIC GROWTH

Valavanis · ECONOMETRICS

Business Cycles and Economic Growth

JAMES S. DUESENBERRY
Professor of Economics, Harvard University

McGraw-Hill Book Company, Inc.
NEW YORK TORONTO LONDON
1958

BUSINESS CYCLES AND ECONOMIC GROWTH

Library of Congress Catalog Card Number 57-13333

V

17980

THE MAPLE PRESS COMPANY, YORK, PA.

For Margaret

Editor's Introduction

For years many teachers of economics and other professional economists have felt the need for a series of books on economic subjects which is not filled by the usual textbook, nor by the highly technical treatise.

This present series, published under the general title, *The Economics Handbook Series*, was planned with these needs in mind. Designed first of all for students, the volumes are useful in the ever-growing field of adult education, and also are of interest to the informed general reader.

The volumes are not long—they give the essentials of the subject matter within the limits of a few hundred pages; they present a distillate of accepted theory and practice, without the detailed approach of the technical treatise. Each volume is a unit, standing on its own.

The authors are scholars, each writing about an economic subject on which he is an authority. In this series the author's first task was not to make important contributions to knowledge—although many of them do—but so to present his subject matter that his work as a scholar will carry its maximum influence outside as well as inside the classroom. The time has come to redress the balance between the energies spent on the creation of new ideas and on their dissemination. Economic ideas are unproductive if they do not spread beyond the world of scholars. Popularizers without technical competence, unqualified textbook writers, and sometimes even charlatans control too large a part of the market for economic ideas.

In the classroom *The Economics Handbook Series* will serve, it is hoped, as brief surveys in one-semester courses, as supplementary reading in introductory courses, and in other courses to which the subject is related.

Professor James Duesenberry has written a book on *Business Cycles and Economic Growth* which will interest all those concerned with the theory and empirical aspects of the problems which used to be dealt with in the study of business cycles. With the great advances, both in theory and fact, Duesenberry's study covers broader issues than those encompassed in the orthodox studies of cycles.

Author of a distinguished book on the consumption function, an experienced teacher both at the University of Michigan and Harvard (where he is now professor of economics, working in quantitative economics, cycles, and money and banking), and with experience in business economics as well, Professor Duesenberry is well qualified to write this important book for the *Handbook Series.*

Professor Duesenberry's book has two aims: (1) to present a framework for the analysis of the effects of changes in the economic structure on the rate of growth of income and on the stability of the growth process and (2) to explain the business cycles of the past and, in particular, to explain why the depression of the 1930s was so much greater than its predecessors.

The first part of the book is devoted to a detailed discussion of what determines investment and consumption. A chapter is devoted to the influences of corporate financing on business investment and another to the relation of competition to investment. There is also a chapter on housing investment and housing cycles.

Duesenberry then uses these results to get a systematic explanation of the growth of demand and the causes of fluctuations. It is argued that past depressions are the result of shocks due to speculative booms and similar factors and not to a repetitive cyclical mechanism.

The book ends with an analysis of wage-price spirals and their relation to the growth of the money supply.

Seymour E. Harris

Preface

Students of business cycles tend to fall into two classes. There are those who assert bold and simple generalizations which lead to elegant business-cycle models. On the other side there are the realists who always point out that the world is much more complicated than any model can possibly be. Since I have a taste for broad generalizations but lack the strength of character to ignore the complications, I have tried to steer a middle course. I have tried here to give a general explanation of the causes of the growth and fluctuation of income. In doing that I have tried to make the explanation complex enough to account for the main facts of the movement of income and its components, but at the same time I have not hesitated to simplify my description of economic behavior whenever I could do so without distorting the picture very much. As a result my book is full of oversimplifications. Yet my description of behavior is not simple enough to make a nice mathematical analysis possible.

That rather unsatisfactory result has just one justification. I have felt for a long time a need for some frame of reference which would make it possible to judge the significance of changes in the structure of the economy (whether produced by poling or otherwise). We live in a world of changing institutions. The forms of business organization, the structure of capital markets, the character of competition, to mention only a few features of the economic structure, have been evolving steadily and at a rapid rate. Do the changes in the structure tend to make income grow more or less rapidly? Do they make the system more or less stable? Some people are very good at answering these questions by intuition, but more pedestrian souls must work out the answers step by step. The theory of income movements given here is intended to provide a basis for analysis of the impact of structural change on the movement of income.

The reader must judge how useful the results are, but it may be worthwhile for me to add a word about the empirical validity of the argument. I have tried to construct a model of the process of income determination which is at least qualitatively consistent with observation. I have taken some pains to test the plausibility of my hypotheses against a wide range of facts. However, I have not presented here anything more than the sketchiest evidence for my hypotheses. The factual material presented here is, as

Pooh-Bah put it, "merely corroborative detail intended to lend artistic verisimilitude to an otherwise bald and unconvincing tale." I have presented a very large number of unverified hypotheses because it seemed worthwhile to present the whole argument without tiding another five years to produce convincing evidence for each proposition. Undoubtedly many pieces of the argument will have to be modified when subjected to really adequate empirical tests, but I have some hope that the main argument has sufficient validity to make its presentation do more good than harm.

It is a pleasure to record my gratitude to the many friends who have helped me and corrected many of my errors. The theory of investment presented here reflects a long series of discussions with John Meyer and Edwin Kuh. I have had a great deal of help from Carl Christ, Carl Kaysen, John Lintner, Wassily Leontief, Merton J. Peck, Guy Orcutt, Paul Samuelson, and Martin Segal.

Much of this book was written during my stay in Cambridge, England, on a Fulbright Fellowship. Richard Stone, Director of the Department of Applied Economics, kindly allowed me to use the Department's research facilities.

Seymour Harris and Robert Dorfman read the manuscript with great care and made many valuable suggestions.

James Duesenberry

Contents

CHAPTER 1

Introduction

The old-fashioned pastime of making business-cycle theories is once more in style. But the theories produced in recent years differ in a fundamental way from their predecessors. The concepts used in modern business-cycle theories are very different from those used before 1930, but perhaps more important than that is the fact the questions which a cycle theory is supposed to answer have changed.

Before 1930, most "orthodox" economists considered that capitalist economies normally tend to operate at or near the full-employment level of real income. Real income could be expected to grow about as fast as capital accumulation, population growth, and the growth of technical knowledge would permit. The idea that aggregate demand might not grow fast enough to absorb the increase in output made possible by the growth of knowledge and resources was considered to be fallacious and paradoxical. The logical basis of that position was the neoclassical general-equilibrium theory. That theory, though static in form, explained growth because it explained the equilibrium rate of capital formation and provided a basis for determining the effect of capital accumulation and population growth on the equilibrium position of the system.

Fluctuations in income and prices were, of course, observed, and they could not be explained in terms of static theory. Neoclassical business-cycle theory explained those fluctuations by introducing lags and rate-of-change variables which are necessarily ruled out of consideration in a static theory. But it was always supposed (on grounds which were not always made clear) that there were forces in the system which would prevent the variables from departing a great deal from their equilibrium positions.

There might have been more skepticism about the validity of the neoclassical position if the facts had not borne it out so well. But, after all, real income did grow consistently for a long time in capitalist countries. There was no secular increase in unemployment, and the fluctuations in real income were very limited.

The factual basis of the neoclassical position was completely undermined by the great depression of the 1930s. The depression could not easily be explained in terms of dynamic qualifications to a static theory of full employment. The experience of the great depression opened the way for systematic examination and attack upon the postulates of neoclassical theory. The upshot of the work of Keynes and his followers was a complete change in our approach to the problem of business cycles. It is no longer considered adequate to regard business cycles as departures from the equilibrium path determined by static general-equilibrium theory and the secular increase in capital, population, and technical knowledge.

Nowadays the task of the cycle theorist is more broadly conceived. The equilibrium theory which explained the adjustment of the secular growth of demand to the secular growth of potential output is no longer generally accepted. It is therefore necessary to explain (1) the processes which account for secular growth in aggregate real demand, (2) the processes by which the secular growth of demand is adjusted to secular growth in potential output (to the extent that such an adjustment takes place), (3) the causes of fluctuations in aggregate demand.

A number of writers have attempted to solve all three problems in terms of a single unified theory. Harrod, Hicks, Goodwin, and Kaldor[1] have all used multiplier-accelerator processes to explain the growth of demand. They then invoke limitations on resource supplies—labor, physical capital, or credit—to explain fluctuations

[1] J. R. Hicks, *A Contribution to the Theory of the Trade Cycle*, Oxford University Press, London, 1950.

N. Kaldor, "A Model of the Trade Cycle," *Economic Journal*, vol. 50, March, 1940; "The Relation of Economic Growth and Cyclical Fluctuations," *Economic Journal*, vol. 64, March, 1954.

R. Harrod, *Toward a Dynamic Economics*, Macmillan & Co., Ltd., London, 1952.

R. M. Goodwin, "The Nonlinear Accelerator and the Persistence of Business Cycles," *Econometrica*, vol. 19, January, 1951.

in income. As a by-product they obtain an explanation of the adjustment of secular growth of demand to secular growth in potential output.

Hansen,[2] by contrast, supposes that the movements of autonomous investment provide the primary explanation of the movements of demand. Autonomous investment is generated by invention, population growth, and the exploitation of frontier areas. Hansen argues that multiplier-accelerator interaction amplifies the income movements caused by fluctuations in autonomous investment.

Those theories are discussed in some detail in the next two chapters. But one feature shared by all of them calls for comment. Each one of the theories mentioned explains the fluctuations of demand by means of a single mechanism. Each author has a different explanation of "business cycles," but each uses only one explanation.

In effect all fluctuations (or, more accurately, all the so-called major fluctuations) in aggregate income are treated alike. That unity of treatment applies not only to the cause of the downswing but also to the cause of the recovery and the continuing movement of income until the next downswing. Everyone, of course, admits that each cycle has its own special features, but those differences are given little emphasis. Mr. Hicks is explicit on the point, but his attitude is typical of the implicit views of the others. Hicks says[3]:

> That the economic systems of modern times are liable to fluctuations of a particular sort, which can properly be called cyclical, is a very obvious inductive generalization from the main facts of economic history. Statistical theorists may warn us that purely random sequences of events (when those events are marked by quantitative characteristics) have a definite tendency to build up into cyclical patterns; and practical statisticians may demonstrate that the facts of economic history can be analyzed to show a bewildering variety of cycles, few of which would leap to the eye unaided by calculation. The plain economist must, I think, conclude that these investigations cancel one another out. The economic history of the last 150 years

[2] Alvin H. Hansen, *Business Cycles and National Income*, W. W. Norton & Company, Inc., New York, 1951.

[3] Hicks, *op. cit.*, p. 1.

organizes itself so easily into a series of 7 to 10 year cycles, with certain interruptions traceable to major wars, that the reality of the cycle seems to him unmistakable.

Even a moderately close examination of the history of American business cycles makes one skeptical of Hicks' position. There are, after all, marked differences in the durations of American depressions and in the amplitudes of the fluctuations in income. It appears to me at any rate that Hicks and the others mentioned above have really been trying to explain a cycle in which all the depressions are like the one which followed 1929. Now, all depressions have some resemblance to one another, but in the respects which are relevant to the cycle theories used by Hicks, Kaldor, Goodwin, and Harrod there are marked differences between the depression of the 1930s and its predecessors. The theories in question explain the lower limit to the decline in income by (1) the physical limits on the rate of disinvestment and (2) the existence of a certain amount of autonomous investment. The recovery is supposed to be due to the effects of capital decumulation and to a rise in autonomous investment.

After 1929, gross investment in fixed capital did indeed fall by almost 75 per cent, and we may legitimately suppose that autonomous factors prevented it from falling to zero.[4] That fits the theory in question well enough. But what of 1920–1921 and 1907? In 1920–1921, the Federal Reserve Board's index of durable-goods production declined by 25 per cent but had recovered to its predepression level within a year of the onset of the depression.[5] There could not have been, in so short a time, any significant amount of capital decumulation. Nor could there have been any very important change in the factors which are supposed to produce changes in autonomous investment. There does not seem to be anything in the theory in question which can explain a quick recovery like that of 1921. (There is, of course, no difficulty in finding plenty of other explanations, but they will imply other causes for the depression.)

In 1907, Frickey's index of durable-goods production fell by 30

[4] *Economic Report of the President*, January, 1956, p. 166.

[5] *Historical Statistics of the United States, 1789–1945*, Washington, 1949, p. 180.

per cent, but once again the index took only a year to recover to its predepression value.[6] The same difficulty, therefore, arises in explaining the recovery of 1907.

One might, of course, say that the depressions of 1907 and 1921 were exceptional, and hope to explain the others by means of a single theory. That would, of course, raise the question why thirty-six years elapsed between the cyclical peak of 1893 and that of 1929.

But the situation is not really any better when we try to apply Hicks' approach to the depressions of the nineteenth century. Consider the American depression of 1873. Frickey's index of durable-goods production declines by 33 per cent from 1872 to 1876 (though by only 5 per cent from 1871 to 1876), then slowly rises by 28 per cent to 1878, and increases another 35 per cent by 1879.[7] Even supposing that capital consumption was 50 per cent of the predepression investment, there was no general decumulation of capital.

According to Hicks' theory, the limit to the downswing in a depression is supposed to be set by the physical limits on capital decumulation and by rising autonomous investment. But there was no capital decumulation. In view of what happened to railroad building in the 1870s, it is hard to believe that there was an increase in autonomous investment between 1873 and 1876. In the same way, rising autonomous investment and the cumulative effects of disinvestment are supposed to produce the recovery. There was no disinvestment, yet railroad building did recover markedly at the end of the depression. However, a recovery of autonomous investment to its former level is not sufficient to produce a recovery in a model like the Hicks' one. In Hicks' model, autonomous investment must rise above the level it reached in the previous boom in order to produce a recovery. That is so because a higher level of autonomous investment is required to offset the adverse effect on induced investment of the decline in income and the continued accumulation of capital.

[6] Edwin Frickey, *Production in the United States, 1860–1914,* Harvard University Press, Cambridge, 1947, p. 64.

[7] *Ibid.*

Similar arguments can be made for the depressions of 1883 and 1893.

We must conclude, then, that, while the theories based on the acceleration principle fit the events of 1929 in some respects, they do not give an adequate account of the earlier major depressions. The factors which limited the downswing and brought about the recovery in the earlier cases did not operate in the same way in 1929. We cannot account for what happened in 1929 in precisely the same way as we account for the earlier depressions. And, of course, the differences between the pre-1929 depressions are almost as marked as the difference between any one of them and the 1929 depression.

No theory can be regarded as really adequate unless it accounts for those differences in one way or another. In view of the complexity of the economic system it is possible to imagine that the observed movements of income could have been produced by a single dynamic mechanism with fixed properties. No one has produced a model which does that, but, even if it were done, the results would be suspicious. We know that the structure of the economic system has been changing rapidly in the last century. Competitive structures and financial arrangements and the forms of business organization have all changed. The rate of population growth has moved over a considerable range. The relative importance of frontier development and of new industries has varied widely. In the face of those considerations it is hard to see how the observed cycles could be explained by a dynamic mechanism with fixed numerical properties.

One way out of this difficulty is to explain cycles in terms of a dynamic mechanism whose qualitative properties are fixed while permitting changes in the numerical elements of the equations which describe the economic system. Then, in qualitative terms, fluctuations in income have a single explanation so long as the numerical elements of the equations lie within certain ranges. That is the type of argument used by Hicks and the other acceleration-principle theorists. They do not require a system with fixed parameters. They require only a system whose parameters are such that

(1) demand tends to grow faster than potential output and (2) the movement of income tends to reverse itself when the rate of growth of income is set at the rate of growth of potential output. As we have seen, the cycle explanations based on the acceleration principle are not really consistent with the pattern of movement of income and investment which has been observed. In addition, the evidence that ceilings have played an important role in bringing on depressions is elusive at best. Finally, it will be shown below that the models used by the acceleration-principle theorists are much too simple. More realistic models will prove to have substantially different dynamic properties.

If it were really necessary, we could continue to seek for a monistic explanation of fluctuations in income. But such an explanation, however elegant, is bound to have a certain implausibility about it. We have argued that the structure of the economic system changes rapidly. It may be that the changes have been limited in such a way that a single qualitative explanation of cycles is possible, but there is no reason why that must have been so. The system might at one time have conformed to the weak acceleration model proposed by Hansen and at another time have bounced off the ceiling in the manner described by Hicks. At other times it might have followed still other patterns. Moreover, even if a single model would suffice to describe the general properties of the system, there is still the problem of shocks. The introduction of ceilings is one way to cause an arbitrary departure of the variables in the system from the path produced by a fixed set of behavior questions, but such arbitrary shifts can be produced in a variety of other ways. The collapse of speculative booms in particular industries, or in the security markets, can have the same effect as a ceiling on the growth of income. So can a cycle of inventory speculation based on a wage-price spiral. Sudden changes in the balance of imports and exports or monetary crises caused by government actions can have similar effects. There is a certain elegance about the use of a built-in shock producer like a ceiling on output, but there does not seem to be any other reason for failing to consider other types of shock.

In view of the irregular pattern of the observed fluctuations in income, the observed changes in the structure of the economic system, and the variety of types of shock to which the system is subject, it does not seem very likely that all depressions can be explained in the same way even in qualitative terms.

In fact, the theme of this book is the idea that capitalist economies can produce depressions of two radically different types. One type of depression occurs when the structure of the economic system is such that capacity tends to rise faster than aggregate demand. In that case, profits and investment decline, and eventually income begins to fall. Recovery occurs after a period of disinvestment or a change in the structure of the economic system. This type of depression is roughly like the one described by Professor Hansen.

In the second case, the structure of the system is such that sustained growth of demand is possible. There is no systematic tendency for capacity to outrun demand. However, depressions may occur because of shocks resulting from the collapse of speculative booms, monetary crises, and a number of other events mentioned above. Naturally, the amplitude and duration of the resulting depressions will vary with the nature of the shock. But the magnitude of the depression resulting from a given shock will also vary with the structure of the economic system and with the position of the variables governing the behavior of investors and consumers at the time of the shock. The initial impact of the shock will usually (but not always) influence investment more than anything else. The resulting change in investment will change the movement of income and produce a set of secondary reactions in consumption and investment behavior. The magnitude of these secondary reactions depends on the structure of the system and its initial position.

If the structure of the system is such that the reduction in income caused by the shock and the secondary reactions takes place rather slowly, income can begin to rise again when the direct effects of the shock are over. However, if income falls below a certain critical level, whose value depends on the structure of the system, recovery will not take place even after the initial impact of the shock has worn off. Income will continue to decline for a long time, and recovery will not take place until capital has been decumulated or the structure of the system changes.

There is also an intermediate case. If the structure of the economic system is such that sustained growth is not possible, income may, nonetheless, grow for a time, given favorable initial conditions. When some adverse shock is administered to such a system, the decline in income will not be arrested when the initial impact of the shock is over. In that case a depression would have occurred without the shock, but the shock makes it occur earlier.

It is argued here that all depressions prior to 1929 were caused by shocks of a variety of types and that income never fell below the critical value mentioned above. The recovery resulted, therefore, from the elimination or reversal of the original cause of the depression. Secondly, it is argued that the evolution of the structure of the American economy in the period between 1900 and the end of World War I tended to reduce the rate of growth of demand and to reduce the stability of the system. The special virulence of the great depression of the 1930s can be explained in two ways. On the one hand, we can believe that the depression would have occurred, even without the boom and collapse of speculation in securities and real estate, because the system was incapable of sustained growth. On the other hand, we can believe that sustained growth was possible but that the system was so unstable that the market crash pushed income below the critical point mentioned above. In any case, it will be argued that the depression which began in 1929 was much worse than its predecessors because of changes in the structure of the economy which affected both the growth potential and the stability of the system.

The argument which leads to those conclusions is worked out by means of a model of the income-determination process, which may be regarded as a generalization of the models based on the interaction of multiplier-accelerator processes with autonomous investment, which have been used by Hansen, Hicks, and others. Hansen's model, employing a weak accelerator, can be regarded as one special case of the model used here. Hicks' model, which is characterized by extreme instability, is another special case. However, the model used here differs from the other models in a number of important respects.

First, it is shown that business investment depends primarily on the level of income, the size of the existing stock of capital, and the level of profits and retained earnings. Profits, in turn, depend largely on the level of income and on the size of the stock of capital, while retained earnings depend on profits and past dividend rates. When we take all those propositions into account, investment is related to income and the size of the existing stock of capital. That is the notion underlying the acceleration principle. But in the model used here, the immediate effects of changes in income on the rate of investment will be much smaller than they would be in a simple accelerator scheme. As a result, the system has more stability than a multiplier-accelerator model. The stability of the system is also increased by the fact that both saving and investment depend on profits, so that both tend to move in the same direction. It is also shown that lags in the reaction of consumption to changes in income tend to increase the stability of the system. In Chapter 7, it is argued that prior to 1914 the characteristics of housing markets contributed to the stability of the system. However, some special features of the situation in the 1920s made residential construction a destabilizing factor.

In Chapter 9, an income generation model is presented. That model may be regarded as a generalized multiplier-accelerator model taking account of the considerations just mentioned. The model does not make any reference to the concept of autonomous investment. The next chapter is devoted to the analysis of growth processes. It is shown that we cannot regard autonomous investment (i.e., investment associated with population growth, new products, and new processes) as a net addition to the investment which would have taken place in the absence of autonomous investment. It is argued that the net increase in total investment resulting from autonomous investment is much smaller than the amount of autonomous investment itself. However, it is argued that during the last part of the nineteenth century autonomous investment played a role of considerable importance.

In Chapter 10, we also consider the effect of a variety of aspects of the economic structure on the ability of the system to maintain steady growth. In Chapter 11, we consider the factors which tend to increase or decrease the stability of the system in terms of the magnitude of the effect of an adverse shock. The characteristic features of various types of shock are also considered.

In Chapter 12, we try to show more concretely why the depression of 1929 was so much worse than its predecessors.

In Chapters 13 and 14, we consider the effects of limitations on the growth of labor supply and money supply. It is shown that the stability properties of the system are such that "ceilings" are not in themselves likely to cause major depressions. They may, however, play an indirect role in bringing on depressions by causing wage-price spirals or monetary crises.

To summarize, then, Chapters 2 and 3 are devoted to destructive criticism of existing cycle theories. In Chapters 4 to 8, various aspects of consumption and investment theory are considered. In Chapter 9, the results are used to produce a model of the income-determination process. The properties of the model are discussed in Chapters 10 to 14. The conclusions reached can be put in the following way: (1) Growth in aggregate demand is produced by a rather complex interaction between a generalized multiplier-accelerator process and certain exogenous factors. (2) At some points, limitations on labor and money supplies have a significant influence on the growth of demand but not in such a way as to produce major depressions. (3) Major depressions have been produced by a variety of different types of "shocks," not by a regular cycle-producing mechanism. (4) The differences in amplitude and duration of depressions are explained partly by differences in the nature of the shocks and partly by historical changes in the structure of the economic system. (5) There may be a certain cyclical element in the situation in the sense that certain types of shock are more likely to occur after long periods of prosperity than immediately after a depression.

CHAPTER 2

Keynes and the Classics

Since the grest depression of the 1930s, economic theory has developed at a very rapid rate. The reader of recent works on business-cycle theory is faced with concepts which were not used at all, or were used in a very different context in neoclassical theory. It sometimes looks as though the theory of income determination had been started over from the beginning with no reference to neoclassical theory.

Concepts like the multiplier and the acceleration principle (at the aggregate level) do not appear in neoclassical theory. Some very similar ideas appear, indeed, in descriptions of the "cumulative process" in neoclassical business-cycle theory. But even in that context, multiplier-accelerator interaction is checked by the operation of neoclassical monetary theory. Cumulative processes of the multiplier-accelerator type came into the picture to a limited extent because it was admitted that price and interest-rate adjustments do not work perfectly and instantaneously. They took over the stage completely when it was asserted that price and interest-rate adjustments are almost completely ineffective in controlling aggregate expenditure.

It can be shown that Keynesian and post-Keynesian monetary theory justifies an approach to the process of income determination which emphasizes the interaction between the rate of investment, the level of aggregate demand, and the size of the capital stock. Neoclassical theory is a special case of that type of interaction in which the dynamic process is checked and controlled by price and interest-rate adjustments. Multiplier-accelerator processes are another special case of the same type of process. But multiplier-accelerator processes use a particularly rigid set of interrelationships and are unchecked by price and interest-rate adjustments.

A more general version of that type of dynamic process takes into account financial limitations (within the firm) on the rate of investment and the correlation between business saving and the rate of investment. Such a process will not be as stable as a neoclassical system, but it will be a great deal more stable than a simple multiplier-accelerator process.

The ideas just sketched out are developed in the next two chapters. This chapter begins with a very brief sketch of the neoclassical theory of the process of capital accumulation. The monetary basis of that theory is then discussed in slightly greater detail. Then the Keynesian and post-Keynesian criticisms of neoclassical monetary theory are outlined. Finally, it is shown that those criticisms require a reformulation of growth theory in terms of a dynamic process emphasizing the interaction of income, investment, and capital stocks.

In the next chapter the acceleration-principle theory of investment and the process of multiplier-accelerator interaction are discussed. The business-cycle theories based on multiplier-accelerator interaction are reviewed. Finally, some modifications in those theories are suggested.

I. NEOCLASSICAL GROWTH THEORY

It is always difficult to discuss neoclassical economics because no one ever set down the postulates and theorems which were generally agreed upon by the analytical economists of the last part of the nineteenth and the first part of the twentieth century (or of any other period, for that matter). Indeed, it is difficult to discuss the views of any one economist. In the nature of the case, economic analysis has a certain *ad hoc* character. The behavior assumptions appropriate to the analysis of one type of problem or situation differ from those appropriate to another. That is so not only because the actual patterns of behavior change from one situation to another but also because what is important about behavior differs from situation to situation. Knowing how to simplify one's description of reality without neglecting anything essential is the most important part of the economist's art. It is therefore unfair to describe anyone's economics wholly in terms of the assumptions he makes in one special situation. It is almost always unfair to say that anyone assumed this or that type of behavior. Few economists of any distinction can be found who did not make a variety of sets of behavior assumptions in different contexts.

They neglected certain variables and relationships in contexts in which the movement of the variables was judged to be small or to have unimportant effects. In a different context the same variables might be given an important role and other variables be neglected. Marshall's apparatus of long runs and short runs and partial equilibria gives explicit recognition to that approach.

In spite of the difficulty of saying precisely what neoclassical theory was, there does seem to be a solid core of postulates and conclusions which were shared by most orthodox economists in the period between the last part of the nineteenth century and 1930.

It is often said that neoclassical theorists were uninterested in growth processes. In fact, however, they had a perfectly coherent notion of the process of capital accumulation and its consequences. They worked it out in a manner which made use of the techniques of static-equilibrium analysis, but the theory was nonetheless implicitly dynamic.

It will be useful to outline that theory very briefly and then give somewhat more detailed attention to the monetary theory underlying the general theory of capital accumulation. We can then show how Keynes and his followers undermined the whole argument by challenging some of its postulates.

In neoclassical theory the rate of capital formation is determined by the same set of processes that determine the rates of production of all commodities. Any change in capital stock or in the rate of capital formation requires adjustments of prices and outputs throughout the economy. If we separate the theory of capital formation from general-equilibrium theory, those adjustments must be taken into account implicitly in the separate capital-formation theory.

To give a clear understanding of the basis of neoclassical capital-formation theory it seems desirable to begin by considering it as a part of general-equilibrium theory. We can suppose that an economy possesses at a given moment a certain stock of various types of capital goods and various types of land. There are a certain number of households which own the capital goods and land. Each household has ordered preferences for combinations of rates of consumption of various types of goods, including saving, and rates of output of labor services. The households rent out land and capital goods, and will supply all the available stock of those items so long as their annual rental is not negative. There is a price for every type of consumers' good and for every type of capital good (we can neglect intermediate products here). There are also prices for labor services and for services of land and of capital goods. These are called wages, rents, and quasi rents, respectively. Finally, there is an interest rate. If the system is in equilibrium, the following conditions are satisfied:

1. Each household consumes goods, saves and supplies labor services at rates which maximize its welfare (in terms of its own preferences) subject to the budget constraint. The budget constraint requires that the value (price times quantity) of consumption per period plus saving equal the value of the services of the factors supplied by the household.

2. Methods of production used for each commodity minimize the cost of production of the commodity at the given prices. The prices include rents of all types, including those for differential ability. Each method of production requires a certain input of each type of factor per unit of output of the product in question.

3. The amount of each factor required in view of the methods of production in use is equal to or less than the amount supplied (if amount required is less than the amount supplied, the price of the factor is zero).

4. The cost of producing each product equals the sum of the factor inputs required times their prices.

5. The cost of each product equals its price.

6. The discounted value (using the interest rate) of the quasi rents (over the service life) of each type of capital good equals the price of that type of capital good. The present value

of quasi rents for a capital good therefore equals the cost of producing it.

7. The value of capital goods produced plus additions to inventory equals saving.

8. The sum of price times quantity produced per period for all commodities equals quantity of money times income velocity of circulation per period.

9. Velocity is a given datum.

In this scheme of things the rate of production of capital goods is determined in such a way as to satisfy the set of conditions just given. There is no unambiguous measure of the stock of capital, but it will generally be true that an increase in the stock of any one type of capital (for given supplies of other types of capital, labor, and given techniques of production and tastes) will lower the quasi rents on all types of capital relative to their costs of production (unless, of course, we choose a type which is important in the production of capital goods or strongly complementary to others). With the indicated exceptions, that is true because the additional capital can be used only by lowering the price of its services until it becomes economical to substitute it for labor or other types of capital. In addition, of course, the price reduction will reduce the relative prices of the goods using the types of capital affected by the change and will lead to an increase in consumption of those goods. If the quasi rents on capital goods fall relative to costs of producing them, the interest rate must fall in order to maintain the equality of discounted quasi rents with costs of production.

It is common to speak loosely of the relation between the interest rate and the stock of capital. Statements in those terms have precise meaning only when the quantities of all types of capital vary in strict proportion to one another. There is no way of adding the different elements in the stock. We can, of course, value the items, but the relative prices of different types of capital goods will change as the amounts available change. Even if they did not, the relation between the interest rate and the value of capital stock would not be invariant to changes in the composition of the capital stock. Nonetheless, it remains true that as all types of capital accumulate, the equilibrium interest rate tends to fall (if population and technique remain fixed).

The neoclassical theorists generally supposed that with a roughly proportional increase in all types of capital the equilibrium interest rate would fall very slowly. That view was, of course, based partly on an expectation that population would continue to increase and partly on an expectation that capital-using techniques would continue to develop. Indeed, Marshall apparently felt that the availability of cheap capital would promote capital-using inventions.[1]

The neoclassical writers seemed to feel that the schedule of marginal efficiency of investment would have a substantial elasticity even if it were not continually shifted upward by technical progress and population growth. They felt that at any one moment there existed a backlog of possibilities for using more capitalistic methods of production. Those possibilities were kept latent by high interest rates. If population did not grow and technique did not change, capital accumulation would lead to a reduction of interest rates and make the exploitation of those latent possibilities economically feasible. It was believed that the annual rate of capital formation was a small fraction of the amount of capital which could be formed before the interest rate would be driven to zero (with fixed population and technique).

The neoclassical conception of the growth process was, so to speak, implicitly dynamic. The static-equilibrium model implies that the rate of capital formation will be a function of the existing stock of capital, size of labor force, consumer preferences, and the set of available techniques. If population, tastes, and techniques are held constant, it is possible to work out a simple dynamic process in which the rate of capital formation depends on the capital stock while the capital stock is obtained by cumulating annual capital formation.[2]

Instead of doing that in a formal way, neoclassical theory worked

[1] A. Marshall, *Principles of Economics*, 8th ed., Macmillan & Co., Ltd., London, 1949, p. 186.

[2] Cf. Robert Solow, "A Contribution to the Theory of Growth," *Quarterly Journal of Economics*, vol. 70, February, 1956.

out the equilibrium position for a given stock of capital and then compared the equilibrium position associated with large stocks of capital with that associated with small stocks.

The conclusion was clear enough. In the absence of technical change and population growth, income would rise with the accumulation of capital, and interest rates would slowly decline. The Austrian theorists seemed to feel that at some low but positive interest rate saving would become zero and the accumulation process would stop. Marshall, on the other hand, seemed to believe that saving would continue even at very low rates of interest.[3] At the same time he seemed to feel that there was no practical limit to the amount of capital which could be used at a sufficiently low but positive interest rate.

It was supposed that growth of population and the development of new techniques would retard the downward movement of the interest rate.

The neoclassical theory of the growth process was therefore put in terms of a theory of moving equilibrium.

II. STABILITY OF EQUILIBRIUM

One can analyze a dynamic process in terms of moving static equilibrium positions provided that each one of those positions is stable. Otherwise a chance disturbance of the system will send it off on a path quite different from the one followed by the succession of equilibrium positions associated with progressively larger amounts of capital.

The stability of equilibrium, in neoclassical theory, was ensured by the use of a number of postulates. Of these the most important was the proposition that the demand for investment is very elastic to the interest rate.

It was supposed that a single year's capital formation would normally be only a small fraction of the addition to capital stocks which could be absorbed before the interest rate was forced to zero (with the existing population and techniques). Ordinarily, then, a

[3] Marshall, *loc. cit.*

year's capital formation would tend to reduce the equilibrium interest rate by only a fraction of a percentage point. That tendency might, of course, be offset by an increase in population or by a change in technique. If for any reason the interest rate did drop by more than a fraction of 1 per cent in one year, the resulting increase in demand for capital would be very large. The elasticity of investment demand to the interest rate was therefore supposed to be very high. In the short run, however, the elasticity of investment demand would be greatly reduced because of the limited supply of the specialized capital used in production of capital goods. In the short run the effect of a fall in interest rates would be offset by a rise in the relative prices of capital goods.

That point is important, however, only when the resources for producing capital goods are fully employed. If the return on investment were to fall for any reason, the specialized resources of the capital-goods industries would become unemployed. A reduction in the interest rate would then cause a movement along the investment-demand schedule until the rate of capital formation was restored to its old level.

On the basis of that argument it was assumed that:

1. The demand for funds for investment is always sufficiently elastic to ensure that there is always a positive interest rate at which full-employment planned investment equals full-employment planned saving.

2. The demand for money to hold (liquidity preference) and the supply of money are sufficiently inelastic to ensure that (*a*) given an appropriate price level, there is no lower limit to the interest rate (except zero) and (*b*) the interest-rate movements required to balance savings and investment can be accomplished with reasonable amounts of inflation or deflation of prices.

3. Prices, and particularly wages, are sufficiently flexible to make the price-level changes required in (2) possible within fairly short periods.

The way in which these assumptions lead to the establishment of an equilibrium level of interest rate and price level and the maintenance of a full-employment income is illustrated in the diagrams in Figure 1. Diagram A shows full-employment planned savings and investment (measured in constant prices) plotted against the interest rate. Two savings functions, S_1 and S_2, are shown. Diagram B shows a liquidity-preference curve in which $k = \frac{M}{\Sigma p_i q_i}$ is plotted against the interest rate. Diagram C shows the relation of the volume of transactions $\Sigma p_i q_i$ to k for a given quantity of money. In

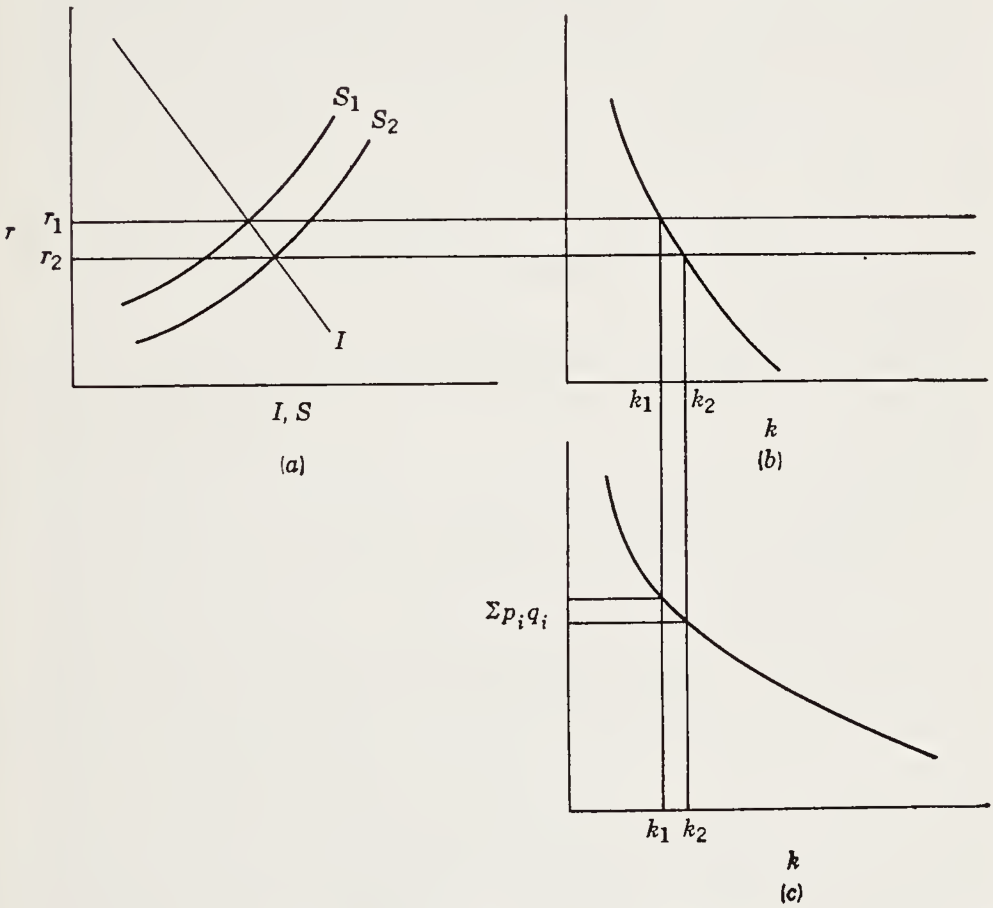

FIGURE 1.

what follows we shall assume for simplicity (though it is not necessary to the argument) that (1) the quantity of money is fixed and (2) savings and investment are independent of real income.

If the savings curve S_1 rules at a given moment, the equilibrium interest rate is set at r_1, and at that interest rate the public will wish to hold money balances equal to k_1 times the volume of money transactions. If the existing quantity of money is M_1, the volume of money transactions must be $\Sigma p_i q_i = \frac{M_1}{k_1}$. The equilibrium relative prices and quantities of transactions necessary for full employment can be taken as given, and the last equation therefore determines the absolute price level.

Now suppose that without any other change the desire to save increases so that the full-employment savings function shifts from S_1 to S_2. From the diagram it is easy to see that the interest rate must fall from r_1 to r_2, k must rise from k_1 to k_2, and $\Sigma p_i q_i$ must fall from $\frac{M_1}{k_1}$ to $\frac{M_1}{k_2}$. With the same "real" situation the Q's and the relative p's must remain constant, so that all prices must ultimately fall in proportion to the rise in k.

How is this adjustment accomplished? We shall suppose that the shift in the S curve takes place slowly, so that speculative effects and cycles of the cobweb type can be neglected. The shift in the S curve causes an excess supply of funds in the security markets, which leads to a fall in the interest rate, an increase in investment, and some check to the tendency for savings to increase. If the liquidity-preference curve were perfectly inelastic, a gradual shift in the savings curve from S_1 to S_2 would result in a gradual fall in the interest rate and a gradual rise in investment and saving. There would be no change in the volume of money transactions. However, if the liquidity-preference curve is somewhat elastic to the interest rate, there are complications. If there is to be a fall in the interest rate, some security holders will prefer to hold cash and some of the planned (Robertsonian) savings will flow into idle balances instead of into investment—planned savings will not equal planned investment. This will lead to a fall in money income. If we hold prices constant for the moment, there will be a fall in real income.

Suppose the savings and investment functions are independent of real income. Then the reduction in real income must be proportional to the increase in k from k_1 to k_2. When that is accomplished, the larger volume of savings will flow into investment. The increased desire to hold balances has been satisfied by the reduction in transactions.

If prices are flexible, the lower level of real income will not persist. Owners of unemployed factors will offer them at lower prices, and all prices will fall. That will cause $\Sigma p_i q_i$ to fall below the equilibrium level and will increase k. Balance holders will begin to buy securities. Interest rates will tend to fall, and planned investment will tend to exceed planned saving. An upward movement in employment and income will begin, while prices continue to fall at a decreasing rate. Ultimately real income will get back to its old level, while interest rates and prices will reach their new lower equilibrium level. If all this takes place slowly (and simultaneously), the movement from one equilibrium to another can be gradual and relatively painless. It is obvious that a similar analysis could be applied to changes in the investment function, the liquidity-preference function, the supply of money, or the equilibrium levels of the quantities exchanged and relative prices.

It is also clear that if sudden changes of appreciable magnitude take place, the adjustments will inevitably overshoot the mark, and cyclical movements in the variables will result. These will be reinforced if there are speculative movements in the liquidity-preference and investment functions. Any number of cycle theories can be generated from this framework by varying assumptions about relative speeds of adjustment of the functions, the types of speculative effects introduced, and emphasis on nonlinearities in the system. Allowance for limited income effects on savings and investment creates still further opportunities for fluctuation in the variables. In effect nearly all of neoclassical cycle theory consists of ringing the changes on the possibilities of the system just indicated. (Structural maladjustment theories, of course, require more microdynamic machinery, but even those theories can be said to work through the kind of mechanism just outlined.)

The introduction of income as a determinant of the rate of saving and of income relative to the capital stock as a determinant of investment brings multiplier-accelerator elements into the theory. But it is essential to the spirit of the neoclassical theories that the effectiveness of those elements (particularly of the accelerator elements) should be limited. That can be done in either of two ways or by any combination of them. On the one hand, a high degree of price flexibility, coupled with high interest elasticity of the difference between savings and investment at each level of income and a low interest elasticity of demand for money, will restrict the operation of any cumulative processes of the multiplier-accelerator type which might get started. On the other hand, if the income elasticity of saving is high and the income elasticity of investment fairly low, the quantitative importance of multiplier-accelerator processes will be low even if they have free rein to operate. (Of course, multiplier processes can never work themselves out fully on the upswing, but it is most important whether their operation is checked by rising interest rates or whether they "collide" with a full-employment ceiling.)

Granting the assumptions discussed above, there is quite a variety of patterns of income which could arise as a result of variations in the responses of the system and of the exogenous factors impinging on it. We need not try to discuss them here. The most important possibilities have been analyzed by Professor Haberler in his admirable work *Prosperity and Depression.*[4] Here we need emphasize only that on the assumptions in question the departures from equilibrium are always limited (unless, of course, there is a continuing sequence of adverse outside forces). The force of any cumulative process leading the system away from equilibrium always tends to exhaust itself, and when it does so, it is succeeded by a reverse process which will suffer the same fate. The economic variables therefore fluctuate about their equilibrium values, and (but for exceptional circumstances) the fluctuations are of tolerable magnitude.

Much of the work of economists in the field of aggregate economics and monetary theory over the last quarter century can be regarded as directed toward the destruction of the assumptions supporting the propositions just made.

[4] G. Haberler, *Prosperity and Depression,* League of Nations, Geneva, 1939.

III. KEYNES' CRITICISM OF NEOCLASSICAL THEORY

Keynes attacked neoclassical theory on a number of grounds. He argued that savings are inelastic to the interest rate but responsive to changes in income. Neoclassical theorists had never (with the possible exception of some Austrian writers) believed very strongly in the elasticity of savings with respect to the interest rate. The relation between saving and income was discussed by Marshall, among others, but in a different context. Keynes emphasized the role of saving as a determinant of aggregate expenditure. In neoclassical theory, aggregate money expenditure was determined by the quantity and velocity of money. "Real" expenditure was determined by the available resources and techniques. The relation between saving and income was important in determining how the pace of capital formation would vary as real income rose and interest rates fell. The possibility of using the propensity to save as a determinant of the level of output was ruled out until other parts of the system were changed.

From a theoretical point of view, Keynes' main attack was based on the theory of liquidity preference. Keynes argued that the ratio of demand for money to hold to transactions is a function of the interest rate. Moreover, the function is of such a nature that it becomes very elastic at some low interest rate. In its simplest form the argument was that a reduction in interest rates would induce holders of securities to try to convert their securities into cash in the expectation that interest rates would later return to their old level. Keynes argued that the interest rate could not fall below a certain level (relative to its historical position) regardless of how great a decline in transactions took place.

If the intersection of the full-employment investment and savings schedules took place at an interest rate below the minimum level of interest rates mentioned above, the equilibrium interest rate could not be reached. No matter how low prices might fall the interest rate would not come down to the required level. Real income would therefore have to fall. The system would finally be brought into "underemployment equilibrium" by a downward shift in the relation of savings to interest rate.

Keynes' equilibrium is only a short-run one. One would expect that with the passage of time investors would give up the idea that the interest rate would return to normal. The interest rate would then fall, and full-employment equilibrium would be restored. (That would be true, of course, only if the saving and investment schedules intersected at a positive rate of interest.)

The real effect of Keynes' argument was to introduce a long lag in the adjustment of interest rates to downward shifts in investment demand. He reinforced that argument by emphasizing the unwillingness of wage earners to take reductions in money wages even when unemployed.

If there is a long lag in the reaction of prices and interest rates to reductions in investment demand, real income can fall by a large amount before those adjustments begin to work. That opens up the possibility that investment demand will fall under the impact of the reduction in income. If it falls far enough, income may not rise even if the interest rate falls to zero.

Keynes did not examine that possibility but stopped with his short-run system.

Some post-Keynesian writers have carried the attack on the neoclassical postulates still farther. Hansen[5] argues that investment demand is very inelastic to the interest rate over the range of rates from zero up to something like 8 per cent. At the same time, Hansen appears to believe that the demand for money has a substantial elasticity to the interest rate at all levels of rates, not merely at very low rates.

If those postulates are accepted, the volume of money expenditures can vary over a wide range.

There is nothing except the response of saving to real income and the stickiness of prices to prevent the volume of expenditures from falling to zero if real investment demand declines below real full-employment savings. On the other hand, if real investment demand

[5] Alvin H. Hansen, *Business Cycles and National Income*, W. W. Norton & Company, Inc., New York, 1951.

rises above real full-employment savings, expenditures can rise until the interest rate becomes high enough, i.e., above 8 per cent, to influence the rate of investment. The monetary mechanism therefore permits the volume of money expenditure to vary over an extremely wide range.

That is an extreme statement of the post-Keynesian position, but there does seem to be wide agreement that the monetary mechanism permits a much wider range of variation in total money expenditure than was believed possible by neoclassical theorists.

A shift in investment demand which generates a volume of expenditure in excess of the full-employment volume at existing prices is not automatically offset by rising prices and interest rates. The prices will rise, but the interest rates need not rise enough to reduce investment to the amount which just produces full employment. The opposite is true for reductions in investment below the full-employment level. It is therefore possible to suppose that real income can vary from the level at which net saving is sufficiently negative to balance the maximum rate of negative investment to the full-employment level.

By contrast, the neoclassical postulates ensured that real income could never vary much from the full-employment level.

IV. POST-KEYNESIAN INVESTMENT THEORY

Neoclassical investment theory was firmly grounded on the general theory of value-and-output determination. As long as the system was at or near its equilibrium position it was safe to say that the marginal efficiency of capital depended on the state of technique, the tastes of households, the size of the labor force, and the existing stock of capital goods.

But the whole argument depended on the postulates which ensured that aggregate money demand would be relatively fixed. Given a fixed money expenditure, all sorts of price adjustments will be effective. If for some reason there is unemployed labor, money wages tend to fall. The marginal-cost curves of firms are shifted down, while the demand curves remain fixed (on the average).

Output and employment rise, and prices fall until the unemployment is absorbed.

It is easy to see that an expansion in the labor force (other things equal) increases the marginal efficiency of investment. The increase in the labor force leads to wage and price reductions and increases quasi rents. Capital-goods prices tend to fall. The prospective returns from investment therefore rise.

In a Keynesian system the consequences of an increase in labor force are problematical. Suppose that an increase in the labor force does reduce wages. Prices of consumers' goods and capital goods will fall, but money expenditure will not remain unchanged. What happens to money expenditure depends on what happens to investment. If the prices of all goods, including capital goods, fall in proportion to wages and if money investment falls in the same proportion, the system is back almost to its starting point. Money expenditure will fall in proportion to the reduction in wages. Both demand curves and cost curves will have moved in the same proportion, and the ratio between capital-goods prices and wages will be unchanged. Keynes argued that the only real change in the system would be a reduction in the interest rates, but his followers regard even that change as unimportant.

That type of argument is, of course, too simple. Capital-goods prices need not fall in proportion to wages. Even if they do, money expenditure will certainly not fall instantly. Therefore the marginal efficiency of investment will be higher than before because capital is being used more intensively. An increase in investment might occur, but there would also be a change in the distribution of income. In practice, of course, there would also be changes in expectations.

We cannot say that wage changes have no effect, but to find what their effects are we have to follow a complicated dynamic process in a step-by-step analysis.

If all that is true or even approximately true, the return on investment cannot be dependent on the balance between the supply of labor and the supply of capital (with given techniques and tastes). In general, it can be argued that so long as there is unemployed labor the prospective return on investment depends on (1) money-wage rates, (2) the prices of capital goods, (3) the level of money expenditure, and (4) the number of units and the age distribution of various types of capital goods in existence (and the rate of change of technique). If money demand increases while wages, prices of capital goods, and stocks of capital goods remain constant, the prospective return on investment will improve. An increase in money demand will lead to both an increase in prices and an increase in output. Higher prices make investment for expansion more attractive. Higher outputs require either the use of equipment inferior to that formerly used or the more intensive utilization (by overtime or extra shifts) of the same equipment In either case the return to be obtained by buying new equipment to avoid high-cost operation will be higher than before. The role played by wages, capital-goods prices, and stocks of capital is obvious.

Those statements will be equally true if there is full employment, but we must add that, if labor is fully employed, the prospective returns from investment for expansion will be lowered because of labor shortages. At the same time the prospective returns from investment in laborsaving devices will be increased if labor is scarce.

All the above statements are perfectly consistent with neoclassical theory, but they are not very relevant, for in neoclassical theory the level of money expenditure depends only on the quantity of money and its velocity. Given that and the composition of the capital stock, wage rates and prices of capital goods as well as the rate of investment and the interest rate are all determined simultaneously by the equilibrium conditions.

The wage rate cannot move without a change in total expenditure or the capital stock, and a change in the wage rate cannot by itself induce a change in total expenditure. Total expenditure (with given capital stock) cannot change without changing wages and the prices of capital goods.

In a Keynesian system, on the other hand, total expenditure can change without changing wages or prices (or at least without changing them much). Moreover, a general change in wage rates can change total expenditure. The neoclassical theory of investment must be regarded as a special case of a more general theory. It holds so long as the neoclassical monetary assumptions hold, but not otherwise. Of course, the same thing is true of the neoclassical theory of growth.

In neoclassical theory, price adjustments were made effective by the inflexibility of aggregate money expenditure. In Keynesian theory, the level of money expenditure becomes the primary subject of analysis instead of being a given datum. Moreover, it is impossible to analyze income changes by static methods.

As we have seen, the level of money expenditure depends in large measure on the rate of investment. But, because it takes time for the multiplier to work out, today's money expenditure depends on the rates of investment in earlier periods. At the same time the rate of investment depends on the level of money income (in the past) and on the size and composition of the capital stock. The rate of investment also depends on the wage rates and prices of capital goods. Finally, the size of the capital stock depends on the sum of the real investment made over a long period. For reasons given earlier, most Keynesian theorists give relatively little attention to the interest rate.

The development of a theory of income determination which emphasizes the dynamic interaction between income, capital stock, and investment was a logical consequence of the Keynesian revolution. Analysis of that type has played an important role in the post-Keynesian literature. Most of the discussion has centered around various versions of multiplier-accelerator interaction processes. Those processes can be regarded as special cases of the interaction between investment, income, and capital stock.

CHAPTER 3

The Acceleration Principle and Multiplier-accelerator Interaction

When there is unemployment and a flexible money supply the prospective return on new investment depends in part on the size of the capital stock relative to the levels of money income and wages. As we have seen, there are very good reasons for supposing that to be so, if we take it as a very broad qualitative proposition. In some of the post-Keynesian literature, however, that basic idea has been pushed a great deal further. The general notion of the relation of returns on investment to the ratio of capital stock to income has been developed into a complete theory of investment—the acceleration-principle theory. Broadly speaking, the acceleration-principle theory makes all (or a large part) of net investment depend on the rate of change of income. The underlying argument depends on some rigid assumptions. Suppose that: (1) At every moment there is an optimum method of production for each product. This optimum method remains the same regardless of the interest rate (over the relevant range of interest rates). (2) The proportions in which different types of products are produced are fixed (with respect to variations in levels of income). (3) Every firm can raise as much capital as it wishes at fixed or approximately fixed rates of interest. (4) The optimum method of production requires a certain amount of capital. (5) Relative prices are fixed with respect to changes in income. (6) In any time period t each firm invests enough to bring its capital stock up to the optimum level for the income ruling in period $t - 1$. Then we can write

$$I_t = \alpha y_{t-1} - K_{t-1}$$

where I_t is net investment in period t, y_{t-1} is income in period $t - 1$, K_{t-1} is capital stock at end of period $t - 1$, α is the capital coefficient or optimum amount of capital per unit of output (income). Now suppose that the capital stock at the end of period $t - 1$ was optimal for income in period $t - 2$. Then

$$K_{t-1} = \alpha y_{t-2}$$

Substituting in the investment equations given above,

$$I_t = \alpha y_{t-1} - \alpha y_{t-2} = \alpha(y_{t-1} - y_{t-2})$$

That is, investment in period t is proportional to the increase in income between period $t - 2$ and period $t - 1$. Then a rigid version of the notion that prospective returns on investment are related to the ratio of capital stock to income is turned into the proposition that the rate of investment depends on the rate of change of income.

Aside from any other defect it may have, the acceleration principle leaves out of account exogenous influences on investment, such as technical change. Many writers repair that defect by lumping all investment not accounted for by the acceleration principle and calling it "autonomous investment." This category is then simply added to investment "induced" by income change. That change can be incorporated into the investment equation given above by writing

$$I_t = \alpha(y_{t-1} - y_{t-2}) + A_t$$

where A_t is autonomous investment. If all investment due to obsolescence is included under autonomous investment, the above equation can be applied to gross investment.

The acceleration principle plus autonomous-investment theory of investment can easily be combined with the Keynesian approach to the theory of consumption.

Keynes emphasized the importance of the link between consumer expenditures and income. A simple form of the "consumption function" makes consumption in period t equal to a certain proportion of income in the preceding period plus a constant. Thus

$$C_t = ay_{t-1} + b$$

where C_t is consumption in period t and y_{t-1} is income in period $t - 1$. If the investment equation is to apply to gross investment, the income terms in both the investment and consumption equations should be gross national product. The difference between gross national product and consumption includes personal and business net saving and depreciation. (Depreciation should, of course, be connected to capital stock, not to income, but that point can be neglected for the moment.) We are neglecting government expenditures, so that the whole gross national product consists of consumption and gross investment. We obtain an equation for aggregate income, then, by adding together the equations for consumption and investment. We have

$$y_t = C_t + I_t = ay_{t-1} + b + \alpha(y_{t-1} - y_{t-2}) + A_t$$

That equation can be generalized by allowing for lags distributed over a number of income periods, but it will be useful to confine ourselves for the moment to the simple two-lag system.

This equation implies that there is an equilibrium level of income given by

$$y_E = \frac{A_t + b}{1 - a}$$

That is, the equilibrium income obtained by applying the multiplier to autonomous investment plus the constant in the consumption function. The equilibrium level of income is derived in the following way. In an equilibrium position, the income is the same in each period, so that $y_t = y_{t-1} = y_{t-2}$. Substituting y_E for y_t, y_{t-1}, and y_{t-2} in the equation above, we have

$$y_E = (a + \alpha)y_E - \alpha y_E + b + A_t$$

which reduces to

$$(1 - a)y_E = b + A_t \qquad \text{or} \qquad y_E = \frac{b + A_t}{1 - a}$$

However, if income is not initially at its equilibrium value, it may not get there for a long time and indeed may never do so.

If the system does not start off at the equilibrium position, it may or may not converge at the equilibrium. For relatively large values of the marginal propensity to consume and capital coefficient the deviations of actual income from the equilibrium will tend to become progressively larger. They will follow an exponential growth path. The path followed by the deviations from equilibrium will be the weighted sum of two simple exponential or compound-interest curves. The weights in the sum will depend on the initial conditions, i.e., on the rate of growth in the first two periods to which the model applies. Regardless of the weights, the system will ultimately converge to the growth path corresponding to the exponential curve having the higher of the two rates of growth. It is important to note that the exponential growth path of the deviations may be in either an upward or a downward direction. The direction of movement depends on the initial conditions of the system as well as on the size of the marginal propensity to consume and the capital coefficient. Let the two exponential rates of growth (in per cent per period) be R_1 and R_2, where R_2 is the larger of the two.[1] If the initial rate of growth $\frac{y_{t-1} - y_{t-2}}{y_{t-1}}$ is greater than R_1, the deviation from equilibrium will move on an increasing path which gets closer and closer to a path increasing at R_2 per cent per period. If the initial rate of growth is below R_1, the deviations will decrease (in algebraic but not absolute value), and income will ultimately tend to decrease at R_2 per cent per period. That does not mean that the income will fall to the equilibrium level and then stop. Income will fall to the equilibrium level and keep on falling until the rate of disinvestment required by the model exceeds the rate of depreciation of capital.

Multiplier-accelerator processes will produce exponential growth when the propensity to consume is around 0.9 and the capital coefficient (with income measured at annual rates) is around 2. The resulting rates of growth tend to be extremely high. They can be somewhat reduced by introducing distributed lags in the reactions of investment and consumption. But it is impossible to obtain reasonable exponential growth rates from a multiplier-accelerator model. One example will suffice. If the marginal propensity to consume is 0.9 and the capital coefficient is 2, the two equilibrium rates of growth will be 13 per cent per year and 77 per cent per year.[2] The

[1] D. Hamburg, *Economic Growth and Instability*, W. W. Norton & Company, Inc., New York, 1956, p. 225 ff.

[2] *Ibid.*

rates would be even higher if a lag period of less than a year were used.

If the value of the capital coefficient is reduced or if the reactions are stretched out over a number of time periods, the multiplier-accelerator process will no longer produce exponential growth. But, if the capital coefficient remains greater than 1, antidamped cycles will result. Income will follow a cyclical path with cycles of progressively greater amplitude. Just as in the exponential case, the movements of income depend in part on the initial conditions. If we start the system off with a rapid upward movement of income, the cycle will start off on its upswing. Conversely, if income is moving upward slowly at the start, the cycle will begin with the downswing.

If we use a capital coefficient of less than 1, the deviations from equilibrium will follow either a damped cycle (in which oscillations become progressively smaller in amplitude) or converge smoothly to zero. In either case the system will tend to reach the stationary equilibrium position.

I. TWO TYPES OF CYCLE THEORY

The four types of income movement just outlined fall into two groups. For relatively high values of the accelerator, multiplier-accelerator processes produce deviations from equilibrium which become progressively larger—either along an exponential path or in cycles of progressively greater amplitude. For relatively low values of the accelerator the deviations from equilibrium tend to become smaller and smaller.

Corresponding to these two sets of possibilities there are two sets of business-cycle theories. Hicks, Goodwin, Harrod, and Kaldor[3]

[3] J. R. Hicks, *A Contribution to the Theory of The Trade Cycle*, Oxford University Press, London, 1950.

N. Kaldor, "A Model of The Trade Cycle," *Economic Journal*, vol. 50, March, 1940; "The Relation of Economic Growth and Cyclical Fluctuations," *Economic Journal*, vol. 64, March, 1954.

R. Harrod, *Toward a Dynamic Economics*, Macmillan, London, 1952.

R. M. Goodwin, "The Nonlinear Accelerator and the Persistence of Business Cycles," *Econometrica*, vol. 19, January, 1951.

The models in question fall into two classes. Kaldor and Goodwin assume that have all constructed theories based on the assumption that the accelerator is large enough and the lags short enough to produce either exponential growth in income or antidamped cycles. Hansen,[4] on the other hand, bases his theory on the assumption that the accelerator is so weak that multiplier-accelerator interaction produces only damped cycles. (The reader may justly wonder why there should be any dispute. We know that the ratio of private capital to income is of the order of magnitude of 2 and has remained in that neighborhood for a long time. Part of the answer lies in the fact that long-distributed lags have the same effect as low capital coefficients. Part of it is due to the fact that the acceleration principle is simply not a very good theory of investment. For the moment, however, we must continue to discuss the cycle theories in the terms in which they have been formulated.)

Hansen's theory is relatively simple. If we suppose that the accelerator is weak, then we must conclude that the long-term movements of income are dominated by movements of the equilibrium position. In Hansen's theory the cause of fluctuations in income is fluctuation in autonomous investment. Multiplier-accelerator interaction merely amplifies the cycle which would be obtained by applying a multiplier to a cyclical movement in autonomous investment. Hansen does not, of course, suppose that the fluctuations in autonomous investment are cyclical in any very precise sense. The fluctuations are due partly to chance, because some innovations are very important and have a "rare-events"

---

the ceiling on output is fixed. If income cannot rise beyond a certain point, investment and income must fall sooner or later (except in very special cases).

Harrod and Hicks, on the other hand, assume that the ceiling on output rises as time passes. The existence of the ceiling will not necessarily cause depressions. Depressions will be brought on by the ceiling only if the rate of increase of potential output is below the lower of the two equilibrium rates of growth produced by the unconstrained multiplier-accelerator process.

The assumption of a fixed ceiling on output or investment is obviously unrealistic. As we shall see, the assumption that the income-generation process is very unstable is also unrealistic.

[4] Alvin H. Hansen, *Business Cycles and National Income*, W. W. Norton & Company, Inc., New York, 1951

distribution. In addition, any disturbance, such as a war, will cause a bunching-up of autonomous investment. Finally, boom conditions favor rapid exploitation of innovations, and depression conditions retard their exploitation. In some periods, then, innovations are used up more rapidly than they are produced, and in others the reverse is true.

Hansen's system can produce secular growth in income by means of a "ratchet effect."[5] Suppose that consumption rises in proportion to income during periods when income is rising to levels above those previously attained. Suppose also that when income begins to fall, consumption declines but not in proportion to income. On the upswing (or rather in that part of it in which income is higher than it has been before), the marginal propensity to consume is high and the constant in the consumption function is zero. On the downswing, the marginal propensity to consume is relatively low and there is a positive constant in the consumption function. If that is true, a cycle in autonomous investment of fixed amplitude can produce an income movement with an upward trend.

The theories based on the assumption that multiplier-accelerator interaction produces explosive movements of income are considerably more complicated. Moreover, there are important differences among the authors of these theories. However, they all use the same basic idea. Suppose that the multiplier-accelerator process tends to produce high exponential rates of growth. Let income initially rise at a rate just above the lower of the two equilibrium rates. (There are some unemployed resources.) Then the rate of increase of income will tend to rise until it reaches the larger equilibrium rate. Suppose that the lower rate is higher than the maximum possible rate of growth of real income. The maximum possible rate of growth of real income is determined by the rate of growth of the labor force and its productivity. When the unemployed resources are all brought into employment, the rate of growth of income is determined by the rate of growth of the labor force and its productivity. The multiplier-accelerator system no longer determines the rate of growth of real income. *Ex hypothesi*, income is now growing at a rate

[5] *Ibid.*

which is below the lower of the equilibrium rates. Once that happens, income will begin to decline. A multiplier-accelerator system of this type turns downward whenever the actual rate of growth falls below the lower of its two equilibrium rates.

Once income begins to go down, it will continue to do so until negative investment becomes as large as possible. Disinvestment will continue for a time, but eventually income will begin to rise again and the cycle will repeat.

A similar argument is possible if the system tends to produce explosive cycles. The argument can also be made to turn on shortages of physical capital goods or credit. Finally, autonomous investment can be introduced into the system in various ways without disturbing the main argument.

That type of theory solves three problems at once. It explains cycles. It explains the secular growth of demand. Finally, it shows why the trend of real demand follows the trend of the potential real output of the system.

However, the whole argument turns on the precise characteristics of the multiplier-accelerator process. If the process did not tend to generate such very high rates of growth and if it were not so unstable, the theory would have no application.

As we have already seen, there are good reasons for supposing that dynamic processes analogous to multiplier-accelerator processes are important to the real world. But it is shown in the next section that there are good reasons for thinking that those processes will not tend to produce extremely high rates of growth of demand and that they are much more stable than simple multiplier-accelerator processes.

II. A CRITIQUE OF THE ACCELERATION PRINCIPLE

It was shown earlier that there are good reasons for supposing that the prospective return on investment will be strongly influenced by the ratio of capital stock to income. It was also shown that the acceleration-principle theory of investment is a special and rigid application of the more general relation between capital stock, income, and returns on investment. Numerous assumptions are required to justify the transition. In particular, it is necessary to assume (1) that the amount of real capital per unit of output which a firm desires is effectively independent of the cost of raising funds for investment or (2) that the cost of raising funds is invariant with respect to the amount raised.[6]

The first assumption has to be evaluated in terms of technological considerations.

There is obviously no technical reason why investment should move in proportion to the rate of increase of income. In the short run, output can always be expanded substantially beyond the level corresponding to what is considered to be the normal or optimal intensity of utilization of equipment. That can be done by using overtime, extra shifts, obsolete equipment, storage to balance seasonal variations, and many other devices. As for the long run, it is clear that more than one method of production is known for almost every product. The different methods are connected with different capital-output ratios. In any case, an increase in demand is not an increase in output. Even if there were only one technique available, the amount of investment generated by an increase in demand would depend on the prospective return required to make firms expand capacity. The size of the prospective return associated with an increase in demand can vary for nontechnical reasons, and so can the return required to induce investment. The acceleration principle clearly cannot rest on a technical argument in a simple sense.

There is, however, a more sophisticated argument. It can be

[6] A. D. Knox, "The Acceleration Principle and the Theory of Investment: A Survey," *Economica*, N.S., vol. 19, 1952.

S. C. Tsiang, "Accelerator, Theory of the Firm, and the Business Cycle," *Quarterly Journal of Economics*, vol. 65, 1951.

D. Hamburg, "The Accelerator in Income Analysis: Comment," *Quarterly Journal of Economics*, vol. 66, 1952.

Hans Niesser, "Critical Notes on the Acceleration Principle," *Quarterly Journal of Economics*, vol. 68, 1954.

J. Tinbergen, "Statistical Evidence on the Acceleration Principle," *Economica*, vol. 5, 1938.

J. S. G. Wilson, "Credit Rationing and the Relevant Rate of Interest," *Economica*, N.S., vol. 21, 1954.

argued that only two or three alternative methods of production are economically interesting at any one point of time. Most of the possibilities are ruled out because they are technically inferior, i.e., use more of all types of resources than the two or three interesting possibilities. Then the same method of production will be used over a wide range of relative factor prices and interest rates.

The significance of the last argument can be greatly exaggerated. It is true that, at any one time, only a very small number of basic processes is likely to be economically interesting in a given industry. But it is also true that (except in a few industries) a wide variation is possible in the amount of capital used per worker in connection with a given basic process. Much or little capital can be devoted to ancillary equipment, such as materials-handling equipment, air conditioning, and office equipment. The amount of space used per worker or machine can be varied considerably (with some return on additional investment in space). A change in the number of shifts worked can have a very large effect on capital intensity. A reduction in the average age of equipment increases the amount of capital invested.

There may be upper and lower limits to the amount of capital used per worker or per unit of output with any given basic process. But in most cases they must be fairly wide, and choice of capital intensity within those limits will depend on relative prices of factors and costs of raising capital. Even small variations in relative prices or costs of raising capital should have some effect on the technique of production and the amount of capital employed.

That amounts to the proposition that, the schedule of marginal efficiency of investment at a given moment is appreciably elastic to the cost of raising funds (over the range of rates of investment which is relevant). One year's investment is only a few per cent of capital stock. If the marginal-efficiency-of-investment schedule is elastic, the rate of investment ought to be sensitive to the cost of raising funds for investment.

It is a commonplace that investment is not very sensitive to the interest rate. But if the interest rate is the cost of raising funds and the marginal-efficiency-of-investment schedule is elastic, investment should be sensitive to the interest rate. That argument may be used to support the view that the marginal-efficiency-of-investment schedule is inelastic. Alternatively, it can be used to support the idea that the interest rate is not the relevant cost of raising funds for investment.

The cost of raising funds for investment is the imputed cost of retaining earnings, borrowing, or selling equities. The imputed cost of retaining earnings and using them for investment is the cost of not paying dividends plus the opportunity cost determined by the available opportunities for financial investment (including debt repayments).

The cost of raising funds from outside sources by borrowing has to include the cash yield given to the suppliers. It also reflects the increase in the amplitude of fluctuations in reported net earnings and dividends which results from taking on fixed charges. In addition, the imputed cost of borrowing has to reflect, to some extent at least, the adverse effect of trading on the equity on the market value of the equity. Finally, it has to reflect the risk that the firm will become a necessitous borrower or go bankrupt. These latter risks, however, must be regarded as fairly small for the larger manufacturing companies which do not have much debt. The effect of these considerations is to raise the imputed cost of borrowing substantially above the cash yield given to the lenders, and to make the cost vary with the ratio of existing debt to profits.

The cost of raising funds by the sale of equities has to reflect the earnings yield given. Earnings yields have been of the order of magnitude of 10 per cent for manufacturing companies.

Whenever a firm needs more capital than the amount available from internal sources, it is faced with a rising supply curve for funds. Indeed, the imputed cost of raising funds must be supposed to begin to rise as soon as the firm wishes to retain more than, say, half of its net profits, for there is some cost of keeping funds away from the stockholders (because of adverse effect on value of equity and pressure from stockholders).

Those considerations help to explain why investment is not very responsive to the interest rate even though the marginal-efficiency-of-investment schedule is elastic. They also show that the acceleration principle cannot be based on the assumption that the cost of raising funds is invariant with respect to the amount raised. The last statement is true whether or not the market supply of loanable funds is very interest-elastic or not.

The acceleration principle will work (in the absence of other disturbing elements) if the marginal-efficiency-of-investment schedule is inelastic and the supply-of-funds schedule elastic. It appears that, on the contrary, the marginal-efficiency-of-investment schedule is appreciably elastic while the supply-of-funds schedule has an inelastic range.

We are not justified, therefore, in using the acceleration principle in a mechanical way. But those arguments do not invalidate the more general proposition that the prospective return on investment depends, *cet. par.*, on the ratio of capital stock to demand. If the cost of funds varies with the amount raised, the actual rate of investment depends on an interaction between the factors influencing the prospective return on investment and those influencing the costs (cash and imputed) of raising funds. There is, so to speak, a market for funds within the firm. However, the demand and supply sides of the "market" for investable funds within the firm have important components in common. As we have already observed, the movements of the demand for a firm's product relative to the size of the firm's stocks of capital goods must be regarded as one of the most important factors governing changes in the position of the marginal-efficiency schedule. The same factors will obviously be important in determining the position and shape of the supply-of-funds schedule for the firm. The supply and demand schedules will not usually move independently of one another. The demand schedule can be moved by changes in technique (or, rather, by changes in the rate of change or type of changes in technique) or by changes in expectations. The supply schedule can be modified by variations in taxes and conditions in the financial markets. But the correlation of the movements of the supply and demand schedules (measured at a given rate of return) is likely to be very high.

The resulting interaction can be illustrated by Figures 2*a* and 2*b*.

In Figure 2*a* the annual rate of profit for a firm is plotted on the vertical axis and the stock of capital on the horizontal axis. The curves marked y_1 and y_2 show the relation of profits to capital stock at two different levels of income, y_1 being the higher of the two, of course. The profits-capital curves have an inverted U shape because there is an optimum stock of capital for any given level of income (we here assume that the firm is a monopoly, so that no account need be taken of actions of other firms). The line at K_0 shows the initial stocks of capital goods of the firm. Figure 2*b* shows the marginal-efficiency-of-investment schedules (marked M_1 and M_2) derived from the total-profits curves for the two levels of income.

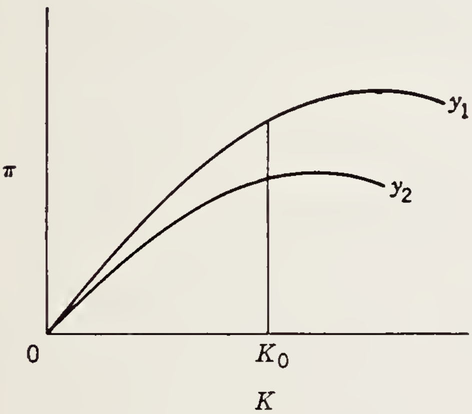

FIGURE 2*a*.

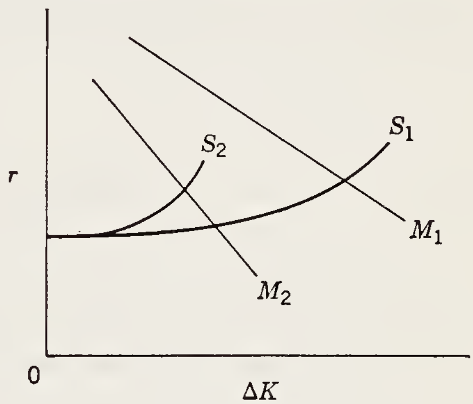

FIGURE 2*b*.

These schedules have, of course, to reflect the durability of the capital as well as its effect on profits. The curves marked S_1 and S_2 are the supply curves for funds generated by the given initial capital stock and the two levels of income (they also reflect, of course, existing debt and money-market conditions).

A fall in demand from y_1 to y_2 changes the relation of total profits to capital stock. Naturally it also shifts the relation of changes in profits to increments in capital stock in the same direction. Since total profits are a major element in determining the supply of funds, and marginal profits a major element in the demand for funds, changes in product demand always tend to move both demand and supply for investable funds in the same direction.

It can be seen from Figure 2 that, when the marginal-efficiency schedule is highly elastic at all levels of income and intersects a rather inelastic supply schedule, the movements of investment will be highly correlated with the movements of the supply schedule. If, on the other hand, a fall in income not only reduces the level of the marginal-efficiency schedule but also reduces its elasticity, the intersection of the marginal-efficiency schedule with the supply schedule may go around the corner of the supply schedule. In that case, of course, changes in the position of the demand schedule will play a much more important role in determining investment than in the other case. Even so there will still be a fairly good correlation between movements of profit and movements of demand.

Since both profits and the marginal efficiency of investment are highly correlated with the ratio of stock of capital to demand (for the individual firm), investment will tend to vary with the ratio of stock of capital to demand.

We can conclude, therefore, that (1) the general notion underlying the acceleration principle is correct, i.e., that an increase in demand relative to capital stock is an important stimulus to investment; but (2) that does not justify the use of any rigid proportionality between increases in income and the rate of investment except under very special circumstances (even if there are no disturbances arising from changes in technique or expectations); (3) the actual rate of investment is primarily determined by an interaction between marginal efficiency of capital on the one side and the supply of money capital to the individual firm on the other. The marginal efficiency of capital is very strongly influenced by the movement of demand relative to real-capital stock, and the supply of money capital to the firm is also strongly influenced by those variables. However, marginal efficiency is also influenced by changes in technique and expectations, and the supply of money capital is influenced by changes in taxes, expectations, and financial-market conditions.

If we disregard the qualifications, we have a generalized version of the acceleration-principle theory of investment.

A. Propensity to Consume and Investment

So far we have been criticizing the acceleration-principle aspect of the multiplier-accelerator-interaction approach. But the multiplier concept is not entirely satisfactory either. In all multiplier-accelerator theories the relationship between saving and aggregate income is supposed to be independent of the factors which influence the rate of investment. Business savings make up about one half of gross private savings. The proportion of aggregate income going into business savings depends on the proportion of income going into profits. We have argued that the proportion of income going into profits is highly correlated with the rate of investment. If income increases in relation to capital stock, investment will increase in relation to income, but the proportion of income saved will also increase.

The propensity to save is not independent of the rate of investment but is, at least in part, dependent on the variables which influence investment.

III. THE STABILITY OF GROWTH PROCESSES

So far we have shown that (1) the marginal-efficiency-of-investment schedule for the firm is likely to have considerable elasticity, (2) the imputed cost of using funds for investment is likely to rise with the rate at which funds are raised, (3) the proportion of income saved is likely to vary in the same direction as the rate of investment. At the same time we have argued that the underlying notions behind multiplier-accelerator theories have considerable validity. It appears, however, that some of the conclusions drawn from the rigid multiplier-accelerator models are wrong or are subject to major qualifications.

In multiplier-accelerator processes of the type used by Hicks, investment moves in proportion to the rate of change of income. If distributed lags are used, it must be said that investment moves in proportion to a weighted average of rates of increase of income over a certain period of time. But the length of the lag period and the size of the proportionality factor are independent of the size of the rate of increase of income. But with an elastic marginal-efficiency schedule and some inelasticity in the supply, things are different. If income rises very rapidly (with business savings fixed), investment will be held back by the rising cost of obtaining funds. The size of the resulting backlog will vary with the rate of increase of income. That is equivalent to saying that the lag between increases in income and the occurrence of investment increases with the rate of increase of income. As a result the cumulative process which tends to produce very high rates of growth of income will be checked. In fact, of course, the size of the backlog will be restricted by the tendency for profits and savings to rise relative to income as the rate of utilization of capital increases. But an increase in the savings ratio also tends to prevent the rate of growth of income from rising. The two factors work together to reduce the strength of the cumulative processes inherent in multiplier-accelerator interaction. It will be shown in Chapter 10 that some other elements in the system have a similar effect.

As we have seen, simple multiplier-accelerator models with large accelerators are very unstable. If income starts to rise, it may continue to do so until checked by an outside factor, or it may go through the upswing of a cycle of increasing amplitude. But if the rate of growth of income falls below a certain limit, income will begin to fall and will decline rapidly.

A model which takes into account the correlation between movements in investment and movements in saving as well as the elasticity in the marginal-efficiency-of-investment schedule will have more stability than a simple multiplier-accelerator model. It is shown in Chapter 11 that a realistic model of the income-generation process will produce only minor cycles in income when its growth is checked by limitations on the labor supply of capital goods or money supply. The ceiling theory of business cycles must be rejected.

IV. AUTONOMOUS INVESTMENT

If we reject the ceiling theories, we are brought back to a consideration of Professor Hansen's theory. Hansen takes the position that multiplier-accelerator processes by themselves are too weak to generate sustained growth in income. On that point we have no real evidence. The observed correlations between the rate of investment and the rate of change of income are very low, but that only shows that the simple rate-of-change version of the accelerator is inadequate. The correlations between profits, stock of capital, and investment are much more satisfactory. That is in accordance with the notion that the rate of investment is limited by financial factors. However, one cannot judge from that evidence that a model based on a generalized version of multiplier-accelerator interaction actually accounts for the growth that has occurred. One is free to imagine that, if there had never been any frontiers or great new industries, there would have been no secular growth in income. It will be argued in Chapter 9 that it is not implausible to suppose that income could grow steadily without a great deal of help from the factors responsible for autonomous investment. However, there can be no doubt that the rapid development of new industries and areas of settlement tends to increase the rate of growth of income. Whether autonomous investment makes the difference between growth and no growth is problematical (since the world changes, it may have done so at some times and not at others), but it is certainly a factor in determining the rate of growth.

But aside from the question of whether Professor Hansen's vision is correct, there are some other defects in his approach. There is no doubt that inventions, new areas of settlement, and so on, have a direct influence on investment. But it is also true that the investment made in a new industry may be offset by lower investment elsewhere. "Autonomous investment" cannot be simply added to "induced investment."

Suppose that a new industry begins to make sales at the expense of other industries almost as soon as it invests in new plant and equipment. Suppose that the industries losing sales would have invested (on a gross basis) had they not lost any sales and that they are just about as capital-intensive as the new industry. In that case a large part of the effect of investment in the new industry will be offset by lower investment elsewhere.

It is shown in Chapter 10 that autonomous investment will have an important net effect on the movement of income when there is a long lag between the investment expenditure and the resulting output. Such a lag may arise either because it takes a long time to build the plant or because economies of scale justify building ahead of demand.

It is also shown that autonomous investment has a net effect on the rate of growth of income when the new industries are more capital-intensive than the ones they displace.

The events which produce what is called autonomous investment play an important role in the income-determination process. But the net amount of investment generated by these events is not as great as the gross amount.

The two types of theories we have been discussing are at opposite poles, and each is based on exaggerated emphasis on some aspect of the situation. The ceiling theories exaggerate the rate of growth that will be produced by the interaction of income, investment, and capital stocks. They also exaggerate the instability of the income movements resulting from that interaction. On the other hand, Hansen's theory exaggerates the net importance of autonomous investment and underestimates the importance of the interaction of income, investment, and capital stock. The first group of theories explains cycles by arguing that the multiplier-accelerator process is too strong to permit steady growth. Hansen explains cycles by alleging that they are too weak.

In what follows, it will be argued that a modified multiplier-accelerator process is capable of producing steady growth in income at moderate rates. Moreover, it will have a good deal of stability. Finally, the process itself will be modified by the occurrence of autonomous investment.

The theory of industrial investment briefly outlined above is developed in Chapters 4, 5, and 6. In Chapter 7 we deal with investment in residential construction and in Chapter 8 with the theory of personal consumption. The results of those chapters are then used to explain the growth and fluctuation of income.

CHAPTER 4

The Marginal Efficiency of Investment

The analysis of business investment is necessarily a very complicated matter. After all, investment decisions must be influenced by every aspect of a firm's activities. We have to deal not only with engineering problems but also with finance, price policy, competitive position, and expectations about the future movements of demand, factor prices, and techniques. Because of the complexity of the problem it is useful to introduce a broad hypothesis (with relatively little content) which will help us break the problem into parts. We shall follow tradition in supposing that the firm makes its investment decisions by choosing those investments whose rate of return over cost (or "efficiency") exceeds the marginal cost of raising funds for investment. Our problem, then, is to analyze the determinants for the level and shape of the marginal-efficiency-of-investment schedule and the marginal-cost-of-funds schedule. The hypothesis is empirically justified at least to the extent that many manufacturing firms do compute the prospective rates of return on investment projects under consideration. However, they often make the computations incorrectly, and we do not know how the payoff periods or cutoff rates of return which they use are related to the cost of funds.

In the traditional treatment of the investment problem the marginal-efficiency-of-investment schedule is supposed to be fairly elastic, and the marginal-cost-of-funds-schedule is supposed to be completely elastic. The rate of investment is then determined by the factors governing the position of the marginal-efficiency schedule and by the level of the interest rate. Since the late 1930s there has been a general tendency to suppose that investment is relatively insensitive to the interest rate. That is, of course, inconsistent with the traditional view of the elasticities of the marginal-efficiency and cost-of-funds schedules.

Two explanations are offered. It is sometimes said that because of uncertainty the excess of the expected rate of return over the interest rate required to justify an investment is very substantial. Suppose a firm requires that an investment have an expected yield of 10 per cent after taxes and interest charges before it can be undertaken. If the tax rate on income is 50 per cent and the interest rate 2 per cent, the expected yield before interest and taxes must be 22 per cent. If the interest rate rises to 4 per cent, the required yield before taxes and interest rises to 24 per cent. A 100 per cent increase in the interest rate raises the required return by only 9 per cent. Even if the elasticity of the marginal-efficiency schedule (in terms of expected returns before taxes) were substantial, investment would appear to be inelastic to the interest rate.

An alternative explanation (not inconsistent with the first) is to suppose that the marginal-efficiency schedule (in terms of expected returns) is inelastic to the rate of return. In that case, changes in interest rates will not have much of an influence even if risk is not so important. As we have already seen, the underlying argument runs to the effect that capital requirements per unit of output are technically determined.

In both arguments it is assumed that the schedule of the marginal cost of funds is elastic. Both arguments can be used to support the acceleration principle (with some adjustment for expectational changes). We know that the acceleration principle does not square with the facts very well. That is not very strong evidence against the hypotheses in question because variations in expectations and uncertainty could account for the failure of the acceleration principle.

But the two hypotheses are inconsistent with some other evidence. In almost every empirical study of investment it is found that variations in profit play an important role in determining the rate of investment. Neither hypothesis implies that. There is also some evidence that investment opportunities generated in prosperous years are backlogged and carried over to less prosperous years. That is also inconsistent with the two hypotheses mentioned above. The evidence on those points is discussed at the beginning of Chapter 5.

In this chapter it will be argued that there are reasons for believing that the marginal efficiency of investment has a substantial elasticity in the long run and a quite high elasticity in the short run. In the next chapter it will be shown that for manufacturing firms the marginal-cost-of-funds schedule has an elastic range and then becomes very inelastic. It is also shown that the marginal-efficiency schedule is likely to intersect the marginal-cost-of-funds schedule in the inelastic range. Finally, the length of the elastic range (in terms of annual rate of investment) is shown to depend largely on the annual flow of profits.

That set of hypotheses will explain the insensitivity of investment to the interest rate. At the same time it explains why profits play such an important role and why "backlogging" of investment sometimes occurs.

In this chapter we shall confine ourselves to the analysis of the marginal-efficiency schedule. Financial factors and the analysis of the interaction of the marginal-efficiency schedule and financial factors are dealt with in Chapter 5.

To simplify the problem we shall deal here with the theory of investment for a simple monopoly. The effect of oligopoly on investment decisions is considered in Chapter 6. In a few places, however, it has been necessary to anticipate the results of Chapter 6.

In section I we consider the determinants of the rate of return on investment for a single-product firm which anticipates with certainty that demand will remain fixed at its present level. It is shown that the rate of return on a given investment depends on (1) the level and shape of the demand curve, (2) the size and age of the firm's existing stock of physical capital, (3) past rates of technical progress, and (4) the prices of capital goods and other input factors.

In section II we consider the effects of uncertainty on investment decisions and the relation between past events and expectation for the future.

We then consider the problem raised by the possibility of integration or diversification. It is shown that those possibilities increase the elasticity of the marginal efficiency of investment to the individual firm.

I. THE MARGINAL EFFICIENCY OF INVESTMENT (STATIC CONDITIONS)

The underlying basis of investment in facilities designed to produce a given set of products is always the existence of some possibility of reducing operating costs by making capital expenditure. Investment to make additional output physically possible is a limiting case. Considerations of uncertainty and of expectations are sources of variations in the prospective returns from capital outlays or in the valuation of those returns, but they are not the basis of the return itself. To get at the problem in its simplest form, it is useful to begin by considering the investment problem of a firm which produces a fixed set of products, has a given set of demand functions, and anticipates with certainty that those functions will not change for a long time. To make the problem still simpler, suppose that the firm has a monopoly and produces a single product. Finally, for the moment suppose that it knows, or thinks it knows, the service life of any equipment that it buys.

A. Cost and Demand Curves and Rates of Return

At any moment the firm has a certain marginal-cost curve which may be assumed to have a rising range at some point (see below for a discussion of shapes of marginal-cost curves). It also has a demand curve and a corresponding marginal-revenue curve. Output and price can be determined in the usual way from the intersection of the marginal-revenue and marginal-cost curves. The firm also has the option of purchasing new equipment which will have the effect of changing its marginal-cost curve. If the equipment is purchased, the firm will change its price and output policy so as to make marginal revenue equal to marginal cost on the new marginal-cost curve. Consider the case of a decision to buy or not to buy a particular set of equipment. In Figure 3, output is plotted on the horizontal axis and dollars on the vertical axis as usual. The line marked *d* is the demand curve, the line marked *MR* is the corresponding marginal-revenue curve, MC_1 is the existing marginal-cost curve, and MC_2 is the potential one which can be obtained by investing. If the firm does not invest, its output will be OF and its price P_1. Its gross profits per year will be equal to the area between the marginal-revenue curve and the marginal-cost curve, i.e., AEC. If the firm makes the investment, its price will be reduced to P_2 and its output increased to OG. Its gross profits will be given by the area between the marginal-revenue curve and the new marginal-cost curve, i.e., the area ADB. Gross profits per year will be increased by the area $EDBC$ between the two marginal-cost curves. The gain in profits comes from two sources: (1) the reduction in cost of producing the output which was being produced before the investment, and (2) the profit on the increase in output which is made worthwhile by the reduction in the level of marginal costs (after allowing for the price reduction necessary to sell the additional output). This second component obviously becomes larger as the elasticity of demand for the product becomes greater.

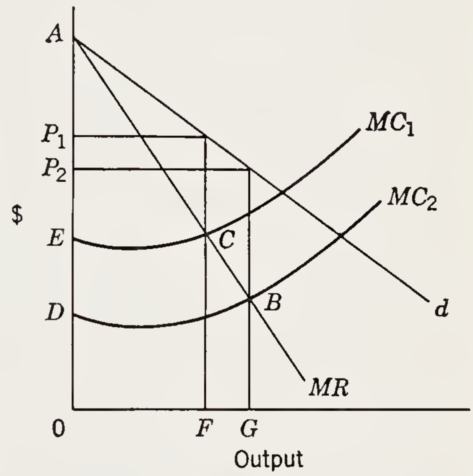

FIGURE 3.

It is not realistic, of course, to suppose that firms adjust their prices in response to every small change in the position of their marginal-cost curves. Nor is it realistic to suppose that they explicitly introduce the elasticity of demand into their calculations of the returns on investment. In a large proportion of manufacturing industries, prices are changed infrequently. Price changes do not fully reflect rising marginal costs when demand expands relative to capacity. Conversely, prices are not reduced when capacity is expanded relative to demand. But when price changes are made, they probably do reflect the cumulative change in the cost situation since the last price change. Since prices are changed, in most cases, at least once a year, there is no great error in taking them into account in individual investment decisions.

But firms do not usually consider explicitly the effect of price changes on sales when they make investment decisions. In a large proportion of cases, firms seem to believe that the market demand for their product is inelastic. At the same time they usually suppose that they cannot increase their market share by reducing price because their rivals will retaliate. It therefore appears appropriate for them to neglect elasticity of demand in making investment decisions. But that is not really the case. It is shown in Chapter 6 that in the long run a firm can take business from its rivals if it can reduce its costs substantially below those of its rivals. When the cost difference becomes sufficiently great, the effect of price reductions will not be wiped out by retaliation because retaliation is impossible. With respect to any sequence of cost-reduction possibilities the firm must take into account the possibility of gaining sales at the expense of other firms if they do not cut costs, or the possibility of losing sales to them if they do. The consideration has to be reflected in the firm's investment policy even though it may be of no importance in individual cases. It may be reflected in the firm's estimate of an acceptable rate of return rather than in its estimate of the rate of return itself, but it has to appear somewhere. For the moment we shall simply assume that firms do take elasticity of demand into account in making investment decisions. The problem of the effect of competitive conditions on investment behavior is considered in detail in Chapter 6.

The annual gain in profits has to be compared with the capital expenditure required to produce it. The "efficiency," or rate of return over cost, of the investment can be computed by finding the rate of discount required to make the present value of the stream of additional profits which the equipment will produce, while it lasts, equal to the capital cost of the equipment. Before we consider the implications of that formulation it may be useful to consider the shape of marginal-cost curves a little more concretely.

B. Shape of Marginal-cost Curves

With a given stock of equipment, marginal costs may be fairly constant over a considerable range of outputs, simply because expansion of output involves expansion of the same operations on similar equipment. At some point, however, it will be impossible to increase output by simply turning on more units of the same type of equipment and working them in the same way. It will be necessary to get additional output by different methods from those used at lower levels of output. The methods available for expanding output beyond the "normal" capacity of the best equipment will vary from industry to industry. In some industries, operating costs increase with the number of hours worked per week. That may be so because extra shifts, overtime, or week-end work must be paid for at bonus rates, or have lower efficiency, or because demand varies seasonally and storage costs are appreciable. In those industries, output can be increased beyond the normal capacity of equipment by working overtime, using extra shifts, or storing output during slack seasons. *Ex hypothesi*, it will cost more to get the extra output than to get the output within normal capacity. In other cases, it may be possible to increase output by closer supervision or by additional maintenance expenditures or closer planning of production flows or by maintaining relatively larger inventories, all of which are costly devices. In still other cases the capacity of a given set of equipment can be increased by subcontracting certain operations at higher cost in order to free equipment for other operations on a larger scale.

In continuous-process industries most of those devices are not available, but in those industries there are often plants which are obsolete and can be brought into use at higher cost. That possibility also exists for other industries, but it is more likely to occur in continuous-process industries because of the tendency for plant and equipment to be so closely connected that major units of equipment cannot be replaced without building a whole new plant. In industries in which plant consists of a roof over the machines it does not pay to keep obsolete equipment because it takes up expensive space.

In general, therefore, we can suppose that a firm's marginal-cost curve has a flat range, after which it rises gradually over a considerable range before becoming vertical.

C. Effect of Investment on Cost Curves

This marginal-cost curve has to be compared with the marginal-cost curve which could be obtained if some new plant or equipment were purchased. The operating costs for performing any operation will generally be lower when new equipment is used than when old equipment is used. In continuous-process industries the method of working any given piece of equipment is fairly rigid. There the effect of buying new equipment is to push the whole of the old marginal-cost curve to the right and insert a block of low-marginal-cost output at the beginning, as in Figure 4. The new equipment may be for any particular process in the industry or for all the processes required to produce the product.

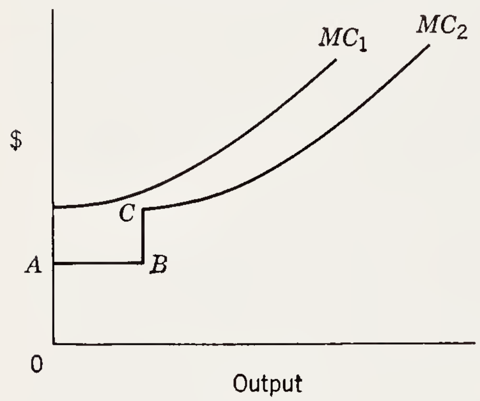

FIGURE 4.

In industries in which equipment can be worked with varying intensity, the purchase of new plants, complete with buildings, not only adds a segment of low-cost output at the beginning of the marginal-cost curve and pushes the old curve to the right, but it also lowers the curve still further in the upper ranges, since plant can be worked less intensively than before.

D. Marginal-efficiency-of-investment Schedule

At any one moment there is only one marginal-cost curve in existence, but the firm may buy any of a variety of new ones by buying various combinations of new equipment. These combinations will involve various numbers of units of each type of equipment and buildings. A rate of return over cost can be computed for each combination of new equipment and buildings in the manner indicated above. After listing the projects in order of rate of return over cost it is possible to construct a marginal-efficiency schedule of the usual type.

However, that approach can be employed only when the effects of the alternative investment projects are independent of one another. If the return on one investment depends on whether or not another investment is made at the same time, we cannot use the marginal-efficiency approach. An example of interdependent returns is the case in which the firm has to decide whether to make piecemeal replacements in an old plant or build a whole new plant. If it builds the new plant, the return on the piecemeal replacements will be reduced because the old plant will not be used so much. On the other hand, if it makes the replacements, the return on building the new plant will be reduced because its cost advantage over the old one will be smaller. This type of problem can be solved readily enough by the present-value method if the cost of raising funds is independent of the amount raised. In that case the firm has three alternatives: (1) new plant, no replacement in old plant; (2) no new plant, replacement in old plant; (3) replacement and new plant. It can choose the alternative which maximizes the present value of future profits. When the cost of raising funds varies with the amount raised, it can do essentially the same thing, but the calculation is more complicated. For simplicity of exposition we shall neglect the problem of the interdependence of returns of different investment projects.

E. Factors Influencing Rates of Return

Once a marginal-efficiency schedule has been computed, it has to be equated to the marginal cost of raising funds. But before we come to that, we need to know a great deal more about the marginal-efficiency schedule itself.

Under the simple conditions postulated above, the level and shape of the marginal-efficiency schedule will depend on (1) the position and shape of the demand curve for the firm's products; (2) the amount and age of the various types of capital equipment which it now has; (3) prices of current inputs; (4) the character of the new equipment which it could buy, including its expected service life and input requirements per unit of output; (5) the prices of new equipment. We shall consider each of these factors in turn.

1. *Effect of Demand*

It seems safe to suppose that upward shifts of the demand curve will, *cet. par.*, increase the prospective return from any amount of new investment. That is readily shown in Figure 5, in which MR_1 and MR_2 represent the marginal-revenue curves corresponding to lower and higher demand curves respectively. The firm's gross profits per year with the low demand and before investment are shown by the area ABC. If it makes the investment required to reduce marginal cost from MC_1 to MC_2, its gross profits will be increased to the area ADF, i.e., by the area $BDFC$. With the increased demand and without any investment its gross profits per year would be BGA. The investment required to reduce marginal costs from MC_1 to MC_2 would raise gross profits to the area DEA, i.e., by the area $BDEG$. The increase in the yield of the investment is therefore the difference between $BDEG$ and $BDFC$, i.e., $CFEG$.

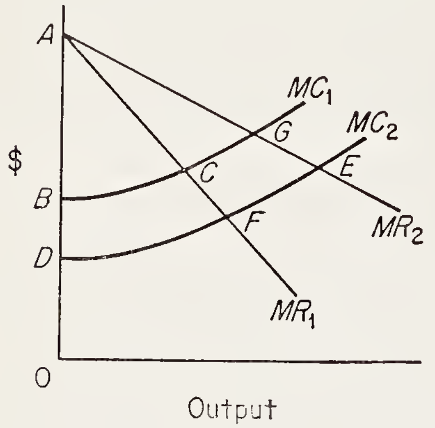

FIGURE 5.

2. *Effect of Stock of Capital*

We can also suppose that, in general, the less physical capital the firm has relative to demand, the greater is the prospective return from additional capital. In Figure 6 four marginal-cost curves are plotted against a single marginal-revenue curve. The marginal-cost curves marked MC_{s1} and MC_{L1} are identical, except for scale, and correspond to the cost situations of two firms, one of which has twice as much equipment of each type and age as the other. The two cost curves marked MC_{s2} and MC_{L2} are then obtained by adding new equipment having

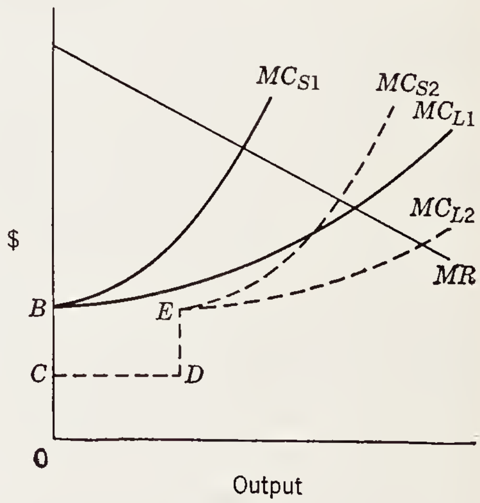

FIGURE 6.

a capacity *CD* to their existing equipment. Both firms will increase their profits by the area *BCDE*, which corresponds to the difference between the operating cost with the new equipment and the operating cost with the best equipment previously in use. But in addition they gain by increasing their output and by reducing the extent to which they have to work old equipment. This gain for the small firm is represented by the area between MC_{s1} and MC_{s2} and the marginal-revenue curve. The corresponding gain for the large-capacity firm is shown by the area between MC_{L1} and MC_{L2} and the marginal-revenue curve. The latter area is obviously smaller than the former. The reason is, of course, that the firm with the larger amount of equipment was not working against increasing costs in the initial situation to the same extent as the other firm. It therefore had less to gain from taking the pressure off its equipment and from lowering price and expanding output.

3. *Effect of Age of Capital*

It can be shown in a similar manner that the profitability of investment varies with the age of the equipment in hand (assuming that there is continuous technical progress or that equipment deteriorates with age).

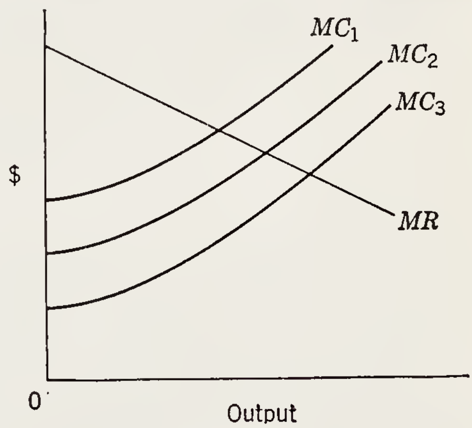

FIGURE 7.

In Figure 7 three marginal-cost curves are shown plotted against a single marginal-revenue curve. MC_1 shows the marginal costs of a firm with relatively old equipment. MC_2 shows the costs of a firm with equipment of the same general type and on the same scale but uniformly newer. MC_3 shows what marginal costs these firms would have if they completely reequipped themselves with the latest equipment. The gain in revenue for the firm with the old equipment is the area between MC_1 and MC_3 cut off by the marginal-revenue curve. The gain in profit for the firm with the newer equipment is the much smaller area between MC_2 and MC_3 cut off by the marginal-revenue curve. The cost curves corresponding to different age distributions of equipment will not usually be parallel to one another, of course, because progress will take place at different rates for different processes.

In general, we cannot draw a sharp distinction between the purchase of equipment for expansion and for modernization or replacement. At any given level of demand the purchase of new equipment almost always results in old equipment's being less intensively used than before, but that is a matter of degree. If some old equipment is already being used for stand-by purposes, the purchase of additional equipment (without an increase in demand) will cause the old stand-by equipment to be junked, while other equipment which was regularly used goes into the stand-by category. In other cases, when old equipment takes up valuable space, new purchases may correspond to retirements. But, in any case, it is not necessary to make any sharp distinction between replacement and expansion for purposes of analysis of investment decisions.

4. *Effect of Technical Progress*

In all the examples given, the rate of return depends on the relative positions of the existing and potential marginal-cost curves. Their relative positions depend in part on the rate at which technical progress reduces factor-input requirements on the newest equipment.

Some forms of technical progress reduce the amount of labor or materials (of the type already in use) required per unit of output. In other cases, technical progress makes it possible to substitute unskilled for skilled labor or to substitute a relatively cheap raw material for a more expensive one. In still other cases, a unit of new equipment may have greater capacity than the old unit. Such a change has much the same effect as a change in the price per unit of the equipment. Of course, any particular improvement in technique may make cost reduction possible in several different ways. In a broad sense, it can be said that the return on investment will rise as the rate of technical development rises, but, since technical improvement reduces costs in so many different ways, there is no precise way in which the rate of technical improvement can be measured.

5. *Effect of Relative Prices*

Investment will also be influenced by factor prices and prices of capital goods. Buying capital goods will almost always reduce the amount of current inputs required to produce a given output. The more expensive the current inputs are, the greater is the increase in profits resulting from the purchase of new equipment, and the higher is the marginal efficiency of investment.

On the other hand, the marginal efficiency of investment is computed by finding a discount rate which equates the increase in profits (before overheads) to the cost of the capital goods which produce those increases. The greater the cost of the capital goods required to produce a given effect on gross profits, the lower will be the marginal efficiency of investment.

The rate of return to be expected for any given amount of investment will increase with the level of demand, the elasticity of demand, the age of existing equipment, the rate of technical progress, and the price of current inputs. It will tend to decrease with the stock of capital and with the price of capital goods.

6. *Elasticity of Marginal-efficiency Schedule*

For given conditions with respect to the factors just mentioned a firm will be faced with a set of investment projects each of which will produce a different rate of return over cost. Some essential replacements or investments required by law (e.g., safety devices) will yield returns which are extremely high because a plant has to be shut down entirely without them. Since many firms use payoff periods (ratio of gross return before taxes and depreciation to capital outlay) of less than three years, it is clear that there are many investments which yield returns of 20 or 30 per cent. At the same time there must be many potential investments which will yield very low returns. Many of those will, of course, never be undertaken. If the firm invests only a small amount, it can confine itself to the investments which promise high yields. If it invests more, it must undertake projects with lower yields. The firm is therefore faced with the traditional downward-sloping marginal-efficiency schedule (subject to the qualifications made earlier about interdependent projects). There seems to be no good reason why the return should center around any particular figure. The amount of investment yielding at least a specified return should increase gradually as the specified return is reduced. The marginal-efficiency schedule ought, therefore, to have a finite slope throughout its range.

Some idea of the magnitude of the short-term elasticity of the marginal-efficiency schedule can be obtained by considering the possibility of speeding up a replacement program. Consider a firm faced with the following situation. A typical piece of new equipment costs $15,000. Improvements in this type of equipment have gone on steadily for a number of years, so that operating costs on equipment produced in year t are $200 below those for equipment produced in year $t - 1$ (for a fixed output). A new piece of equipment, therefore, produces a given output at an operating cost which is $3,000 below the operating cost on a piece of equipment 15 years old, $2,800 below the operating cost on a piece of equipment 14 years old, and so on. Suppose that the firm normally uses a 5-year payoff period as a basis for making replacement decisions. In that case it normally replaces equipment when it is 15 years old.

Suppose that for some reason funds are scarce. For example, demand has increased rapidly, so that the firm wishes to build a new plant. Retained earnings plus depreciation plus an amount of borrowing sufficient to maintain some normal ratio of debt to earnings are insufficient to finance normal replacement plus the expansion.

Before the firm changes its borrowing policy, it will scrutinize applications for funds more closely than before and will in effect reduce the required payoff period. If the payoff period is reduced from 5 to 4.7 years, the firm will delay its whole replacement program by a year. A year's replacement of equipment is a substantial portion of annual investment. Therefore an increase of 6 per cent in the required rate of return (not 6 percentage points but an increase in the required return equal to 6 per cent of the old return) reduces investment by perhaps 20 per cent.

The same argument holds in reverse for decreases in required rates of return. The short-run elasticity of investment must, therefore, be very high. The long-run elasticity is, of course much lower.

II. EXPECTATIONS AND UNCERTAINTY

Variations in expectations above the future values of economic variables and variations in the degree of certainty with which those expectations are held play a central role in the determination of the rate of investment. They must do so, in the nature of the case, because the outcome of an investment decision depends entirely on the values of the relevant variables during the years after the decision is made. The values of the economic variables at the moment of the decision are significant only because they influence the decision makers' expectations about the future. It is probably true, however, that current and recent past conditions are given greater importance than they deserve in determining expectations about the future.

In the preceding section it was assumed that firms expect conditions to remain as they are for an indefinite period. That assumption was made only for expositional convenience. We now have to consider what difference it makes if a firm's management thinks that conditions in the future may differ from those now obtaining.

The discussion of the role of uncertainty has to be carried on at two levels. Starting from the general position that we have some idea as to what we mean by uncertainty and expectation, we can try to formulate hypotheses about the effect of uncertainty and expectations on investment. But such hypotheses are of no use unless they can be put into operational terms (although it is not necessary to start with operational definitions from the beginning). If we can give no operational measures of expectations, uncertainties, or risk preferences which are independent of observations on the behavior they explain, we can produce nothing but tautologies. Ultimately, then, it is necessary to make some statements about the relation between measured expectations, uncertainties, and risk preferences—the measures being those obtained from questionnaires or movements of variables which are believed to govern the expectations and investment behavior. It is therefore necessary to test jointly a hypothesis on the influence of expectations on investment behavior and a hypothesis that certain observable variables are correlated with the expectations in question. Indeed one can go farther. A useful hypothesis requires that expectations be explained as well as measured. If they can be measured but appear to be unrelated to other economic variables, nothing has been added to the explanation of investment. We might as well put a random variable into the explanation of investment and let it go at that. It is, of course, a step toward a useful hypothesis to measure expectations and show the relation between expectations and investment. The explanation of expectations can come later if necessary.

A. Effect of Predictable Demand Changes

We can move one step nearer to reality by considering a firm which can predict changes in future demand with certainty. Firms with divisible equipment will be interested in three aspects of the movements in demand: (1) in the average level over a period of, say, ten to fifteen years, (2) in the frequency distribution of deviations from the average without regard to timing, (3) in the detailed timing of demand movements during the next two or three years.

1. *Average Level of Demand*

At a first approximation, the average level of demand over the life of the equipment in which the firm might invest plays the same role as the fixed level of demand considered in the last section. Other things equal, the higher the average expected level of demand the higher will be the expected return from any investment. Although many investments have a life much longer than ten or fifteen years, it is not usually necessary to forecast for any longer period. In a realistic situation in which full weight is given to uncertainty, risk discounts on prospective future returns are so high that the present value of an investment is not much influenced by the firm's expectations about events in the very distant future, whatever they may be. Although we are considering the certainty case, there is no point in departing any farther than necessary from realistic conditions in which uncertainty plays a leading role.

2. *Variations in Demand*

The frequency distribution of deviations in demand from its average value enters the picture because the occurrence of fluctuations in demand influences the type of equipment which the firm requires. When demand is steady, the firm requires a type of equipment which is optimal for a single level of output. When demand varies, the type of equipment which is most suitable is not the optimal type for any single level of output.[1] The firm then requires a type of equipment which leads to an average total-cost curve which is relatively flat on both sides of the expected average output, even if such a curve has a somewhat higher minimum cost (at the expected average output in the case of a symmetrical distribution of levels of demand) than a plant which is optimal for the expected average output. The type of plant which is most suitable for a fluctuating demand is likely to be somewhat less capitalistic than the type of plant most suitable to a steady output, and for this reason fluctuations in demand tend to reduce investment. The importance of flexibility will be considerably reduced if the firm expects an upward trend in demand.

Fluctuation in demand is important in a quite different way. It is bound to produce fluctuations in reported earnings and in the flow of funds available for dividends. Those fluctuations will, of course, be greater the greater the amount of fixed charges the firm has to bear for interest and depreciation. In the certainty case there is no danger of the firm's getting into difficulty on that account, since it can always store funds to meet dividends and interest charges. Even so, firms dislike fluctuations in reported net and

[1] George, J. Stigler, "Production and Distribution in the Short Run," *Journal of Political Economy*, vol. 47, pp. 305–327, 1939.

the payment of dividends in excess of earnings.[2] They will therefore be willing to take a somewhat lower average profit to avoid fluctuation in earnings. Thus fluctuations in output will tend to lower the marginal efficiency of investments involving large additions to depreciation charges. The same thing would be true of those involving additional interest charges. But because interest charges depend on the method of finance rather than on the character of the investment, it will be more useful to attach the interest aspect of fluctuations in output to the analysis of the supply schedule for funds. It should be noted, however, that fluctuations in the prices of raw materials may lead to greater investment than would stable prices, because firms must have equipment which can use more than one type of raw material. For example, electric utilities often have heating plants which can use oil as well as gas, or oil as well as coal.

3. *Demand in Near Future*

Unless the fluctuations in output are all expected to occur at the beginning or at the end of the life of the investment in question, their detailed timing will not be of much interest in determining the degree of flexibility which a firm should have. But firms will be interested in the detailed timing of demand movements in the first two or three years after the decision date. The expected average level of demand may be high enough to make a given investment profitable. But if it is expected that there will be a short depression immediately after the investment is made, it may be still more profitable to wait until the depression is over before making the investment. That will be true even if the depression has been allowed for in estimating the expected average level of demand.

Our conclusions in the last section have to be modified in the following way to take account of expectations of changes in the level of demand. The marginal efficiency of capital will be lower than the one calculated on the assumption that the current level of demand will persist under the following conditions: (*a*) If the cur-

[2] That proposition has nothing to do with risk preferences. The argument here has to do with the time pattern of the returns rather than with uncertainty as to whether they will be received or not.

rent level of demand is believed to exceed the average future level of demand because the current situation is regarded as a temporary boom due to a war or some other special factor. (*b*) If the current level of demand is judged to be above the average future level of demand because a long depression is expected sometime during the next ten or fifteen years. This consideration is, of course, of less importance to firms in growing industries whose demand will not fall much in a general depression. (*c*) If the current level of demand is not far from the expected average level of demand, but a short depression is expected in the immediate future. (*d*) If large fluctuations in demand are expected, and they lead to the use of a type of plant having a lower capital intensity than would be used with a steady demand.

As long as we consider decisions relating to the purchase of perfectly divisible equipment, the possibility that the future average level of demand will be above the current level is not of much interest. The firm has no need to buy any units of equipment whose employment is not profitable at the current level of demand. If the demand improves later on, additional equipment can be purchased.

4. *Effect of Indivisibilities*

The case is different for firms using indivisible plant or equipment, i.e., firms which use plant or equipment units having a capacity which is a substantial fraction (say, 10 per cent or more) of the firm's output. A large proportion of industrial firms fall into that category. When the output of one of those firms becomes large relative to the capacity of its existing plant, its costs will begin to rise. But, if no further increase in output is anticipated, it may not pay to buy large units of plant or equipment because not much output will have to be produced at high cost. The high rate of utilization will have an effect on investment because it will interact with technical progress. The firm will be led by the high rate of utilization to buy new equipment sooner than it otherwise would because the cost savings to be had from the purchase of new equipment can be added to those to be had from reduction in intensity of utilization. (That implies, of course, that some of the old equipment will be kept in service and that no clear distinction between replacement and expansion can be drawn.) When output is expected to rise in the future, the rate of return over cost is obviously higher for a large investment than it would be if output were expected to be constant. In the latter case, all the elements in the series of returns to be discounted will be the same. In the rising-output case, the early elements will be the same as in the constant-output case and the later elements will be larger, so that the marginal efficiency will be larger. The rate of utilization at which a firm finds it worthwhile to buy a large unit of equipment will therefore be lower when demand is expected to rise than when it is expected to remain constant.

We can conclude, then, that the marginal efficiency of investment in large indivisible units of equipment will be raised above the marginal efficiency calculated on the assumption of the persistence of the current level of demand when it is expected that the future average level of demand will exceed the current level. That will be the case (*a*) when the firm regards the current level of demand as abnormally low and expects a recovery to normal or (*b*) when the firm expects output to rise trendwise in the future (and the expectation is not offset by the possibility of a long depression).

It is possible, of course, for the firm to expect the average level of future demand to be higher than the current level and still expect that things will be worse before they get better. In that case it can only be said that the two considerations pull in opposite directions.

To summarize, the marginal efficiency of capital will be below the value estimated on the assumption of constant demand if average demand in the future is expected to fall below the current level, and above it if average demand in the future is expected to be above the current level provided that indivisible equipment is used. Marginal efficiency will also be reduced by expectations of a decline in output in the near future.

B. Uncertainty

There is some question, first of all, about the way in which expectations subject to uncertainty ought to be described. They are often described in terms of probability distributions, but that formulation has been objected to.[3] It has sometimes been argued that probability distributions cannot be used in describing expectations because the decisions to be made on the basis of those expectations are unique. In that case, knowledge of the expected value of the outcome of a decision is not very significant. That, however, is really an objection to one kind of use of a probabilistic description rather than to the description itself. The statement that the chance of a head on a single toss of a coin is ½ is a perfectly adequate description of the expectations of a man deciding whether to take a bet on the outcome of a coin toss. It does not follow that the man will choose the alternative which maximizes the expected value of the outcome.[4]

Moreover, most investment decisions are not in any significant sense unique. It is true, of course, that no investment decision is exactly like any other. But if numerical probabilities can be attached to alternative outcomes measured in dollars for each one, there can be no objection to compounding the probabilities. The fact that one investment involves a machine and another a building is no more significant than the fact that one pair of dice is white and another one red. The real objection to compounding probabilities arises from the possibility of going bankrupt. If firms had unlimited credit resources, they would maximize profits by making all investments which are better than fair bets (i.e., where the expected discounted value of the returns exceeds the amount invested). But that is not a correct policy if borrowing power is limited. Even if one is concerned only about maximizing the expected value of one's future income, it does not pay to take all bets which are better than fair if

[3] A probability distribution of next year's sales is a list of numerical probabilities (as estimated by the person whose expectations are in question) for each possible level of next year's sales. For example, a firm's expectations might be described by saying that the management thinks that there is only one chance in ten that sales next year will exceed sales this year by over 10 per cent, that there are two chances in ten that next year's sales will exceed this year's by more than 5 per cent and less than 10 per cent, two chances in ten that they will exceed this year's by less than 5 per cent, and that similar probabilities hold for decreases in sales. The sum of the probabilities in the distribution must be 1.

[4] Kenneth J. Arrow, "Alternative Approaches to the Theory of Choice in Risk Taking Situations," *Econometrica*, vol. 19, 1951.

doing so involves an appreciable chance of running out of funds, and being unable to play at all. In fact, of course, managements are concerned about bankruptcy for reasons unconnected with the maximization of expected profits. If a policy which would maximize expected profits involved taking a 10 per cent chance of going bankrupt, no management would follow that policy. Perhaps the simplest rule of behavior is to say that a firm ought to come as near to maximizing expected profits as is consistent with avoiding more than an arbitrarily small chance of bankruptcy. It is impossible to stay in business and avoid any chance of bankruptcy, but it is often possible to follow policies which reduce the chance of bankruptcy to negligible levels.

Of course, firms are concerned not only with expected average profits; they are concerned also with the time pattern of profits. In general, they will be willing to accept a reduction in the expected value of average profits if by doing so they can reduce the chance that earnings will fluctuate over a wide range.

Another weakness of the probability approach arises from the fact that the prediction of economic events is not really analogous to the prediction of card drawings or dice games. It is true that economic events are influenced by events which can be regarded as random (i.e., distributed in a manner analogous to the results of card drawings), though they are usually serially correlated. If we knew the structure of the economic system, the prediction of economic events would be analogous to the prediction of the results of card drawings or dice games. But we can only infer the structure from the past events which it has produced. These in turn were influenced by random factors. To a large extent, then, the probability concept appropriate to the investment-decision problem is that of fiducial probability.[5] If one is trying to make a good score as a forecaster and deals with a number of different structures, he can use fiducial probabilities in the same way as ordinary probabilities. But that is

[5] For these purposes a fiducial probability may be defined as a measure of the frequency with which a forecaster will be correct if he places limits of his forecasts according to some given standard. It is therefore a measure on the confidence which can be placed in a forecast.

not the case when one makes a number of decisions all of whose outcomes turn on the nature of a single structure. Since the economic structure changes slowly, the outcomes of all the decisions made over a considerable period will depend on the nature of a single structure. Fiducial probabilities cannot be compounded in that case.

That is true whether a decision maker projects past cycle and trend patterns into the future, accepts an academic business-cycle theory, or weighs the various arguments of professional forecasters. Each of these notions implicitly or explicitly embodies a hypothesis about the structure of the economic system and calls upon some empirical evidence in support of its argument. No one believes wholeheartedly in any one of them, and nearly everyone is aware of alternative hypotheses or approaches to forecasting. When they disagree, all but one (at least) must be wrong.

In these circumstances there is no real meaning to a probabilistic statement of the situation if probability is interpreted in terms of frequency distributions. We may say that one event is more likely to occur than another, but that is an expression of confidence in the plausibility of one hypothesis over another. It is not equivalent to the statement that one head and one tail is a more likely outcome of two coin tosses than either two heads or two tails.

It has been suggested[6] that in these circumstances rational behavior consists in choosing that course of action which minimizes the maximum value (on any hypothesis) of the difference between the gains from the action on a given hypothesis and the largest possible gains from any action on that hypothesis. This argument is based on the assumption that there is complete ignorance of the true hypothesis and that only one decision is to be made. It is true, of course, that, since we do not know what is going to happen, we are, in a sense, completely ignorant of the future except to the extent that only certain things are believed to be possible. That implies that we must be prepared to see our judgment as to the most likely event (in the confidence sense) go wrong. Any rational man therefore tries to choose a course of action which will not do too

[6] See Arrow, *op. cit.*

badly in any event rather than one which will do very well if his judgment is correct and very badly if it is not.

But that does not mean that decision makers do not give any weight to their judgment as to what will happen. A business firm whose management views all possible contingencies as equally likely and makes its decisions on the basis of "minimaxing the regret" is very likely to sell out the assets for what they will bring. A set of contingencies which will result in the loss of all the firm's capital is almost always possible, and those contingencies will usually dominate any minimax calculation. If firms stay in business, their managements must give greater weight to what they believe to be plausible hypotheses than to hypotheses which they regard as possible but implausible. If they do so, they will have to weigh the various possible outcomes from a given action by a "confidence factor," which will have to be treated analogously to a probability. Such probabilities can be compounded for decisions depending on independent hypotheses. That amounts only to saying that a decision maker will suppose that, while he may be wrong some of the time, he will not be wrong all the time.

There are various ways of formulating the decision problem, but it is easiest to connect the discussion of decisions under uncertainty to the earlier discussions if we suppose that the firm first formulates the policy which is optimal for the forecast which it considers most plausible, and then modifies it to take account of the possibility of that judgment's being wrong in various ways.

Since we have not yet come to the supply-of-funds problem, we are discussing not the effect of uncertainty on complete decisions but the effect of uncertainty on the marginal efficiency of investment. The meaning of the marginal efficiency of investment is clear enough under conditions of certainty. When the future is uncertain, we have to define the rate of return associated with a given increase in the amount to be invested as the rate of return on a certainty of equivalent amount, e.g., a government bond which the firm would be just willing to take rather than make the investment in question. We can regard the discussion of the relation between investment and expectations for the certainty case as applying to the relation between investment and the most plausible forecast in the uncertainty case. The problem is to determine how recognition of other possibilities will affect investment decisions.

1. *Decisions under Uncertain Conditions*

Obviously the final decision will be a compromise among the optimal policies for different hypotheses and will not be optimal for any one of them. The firm will have to modify the policy most suited to the "best" forecast. In doing so it must balance the gain from a departure from the policy which is optimal for the best forecast when that forecast is wrong with the loss from the departure when that forecast is right. The calculation will therefore have to take account of the degree of plausibility of alternative hypotheses and the extent to which the firm attaches a greater importance to avoiding a given risk of loss than to obtaining a given chance of gain.

The discussion can be made more concrete if we again consider the effects of differences in uncertainty with respect to (*a*) average level of demand, (*b*) range of fluctuations, (*c*) near term movements in detail. With respect to the average level of demand we can consider three cases: (1) a probability distribution which is symmetrical about the mode (or best estimate), (2) and (3) probability distributions skewed to the right or left of the mode.

In the first case, the firm will wish to have a plant with a greater degree of flexibility, i.e., with a flatter average total-cost curve even if its minimum is higher than it would wish to have if it held its best forecast with certainty. If it has a distaste for risk, it will weigh downward deviations from its forecast more heavily than upward ones.

Some flexibility can be achieved by using more rather than less divisible equipment. Additional flexibility can be obtained by keeping obsolete plant and equipment in serviceable condition instead of scrapping it. If the firm wishes to avoid risk, it can build less plant than it would build for the best forecast, and keep obsolete equipment in working condition, using it a small proportion of the time if its forecast works out, a large proportion if its forecast is too pessimistic, and not at all if its forecast is too optimistic. If its forecast is too pessimistic, it can build additional plant later on if it wishes. The upshot is that, even when the probability distribution is symmetrical about the best forecast, the firm is likely to build less plant under conditions of uncertainty than it would under conditions of certainty.

It seems obvious that the downward bias of investment will be still greater if the firm thinks it very unlikely that demand will exceed its best forecast, but considers it quite likely that demand will fall short of its forecast. This may be the case when there are some signs of a depression. A firm may believe that a depression is unlikely, though possible, and may also think that an increase in demand is out of the question.

When the distribution is skewed in the other direction, the firm can be supposed to invest at least as much as it would in the certainty case and more if it has indivisible equipment. (When it is a question of divisible equipment, the firm does not lose much by waiting to see how things turn out, so that the effect of the upward bias is weakened.)

For any given modal expectations about future average levels of demand, then, investment will tend to be lower than in the case in which the best estimate of future demand is held with certainty if the probability distribution is symmetrical about the mode or skewed downward, and investment will tend to be higher if the distribution is skewed to the right and equipment is indivisible.

Uncertainty as to the range of year-to-year variation in demand will reduce the rate of return imputed to any given amount of investment in two ways: (1) It will increase the advantages of flexibility and (2) it will increase the chances that the firm will sustain very large losses in particular years and therefore make the firm less willing to take on additional fixed charges.

When we turn to the question of detailed timing, some new possibilities appear. Even if a firm considers that the long-term prospects justify building new plant, it still has to decide whether it should do so immediately or wait a bit. It was argued above that in the certainty case an expectation of a short-term decline in demand would cause a firm to delay an investment project which was justified both by the current and longer-term prospects.

When a firm is in doubt as to the situation in the immediate future, the situation is more complicated. We might suppose that the firm has only two alternatives: (*a*) to build the plant which would be optimal in view of its long-term expectations or (*b*) to continue operating with the existing plant. Then the firm will simply have to weigh the profits resulting from each action by the likelihood of each outcome and choose. In most cases, however, some third possibilities will be available which will permit the firm to steer a middle course. It may, for instance, be able to subcontract part of its operations or to use secondhand equipment and lease space to expand output. In those cases its position will always be intermediate between the two more extreme policies. If demand proves to be high, it will have higher costs than if it had built new plant; but if output proves to be low, it will have put much less into unneeded expansion.[7]

The object of the middle-of-the-road policy just mentioned is to avoid making an investment commitment. It seems reasonable to suppose that, when it appears at all likely that demand will fall below present levels, the marginal efficiency of a given amount of investment will be reduced sharply. A fifty-fifty chance of a depression in the next year or two might reduce the marginal efficiency of investment almost as much as certainty of such a depression.

Such a large reduction is not due to the effect of the depression on the average returns from an investment over its life. It is due to the large increase in returns which can be obtained by waiting. A firm seldom has to decide whether to invest now or never. It usually has to decide whether to invest now or wait a year and then reconsider the question.

C. Formation of Expectations

A firm's expectations about future demand depend in part on factors which condition its expectation about the movement of gross national product or "general business conditions" and in part on factors conditioning its view of the relation of the movement of the firm's demand to the movement of general conditions. Let

[7] Cf. Albert G. Hart, *Anticipations, Uncertainty and Dynamic Planning*, Augustus M. Kelley, Inc., New York, 1951.

us consider first the factors influencing expectations about the future of aggregate demand.

It is a plausible hypothesis, and one for which there is some evidence,[8,9] that all demand forecasts are very strongly influenced by present levels of demand. No one knows very much about the future, but we do know where we are. There is a fairly strong tendency for most of us to expect the future to be like the present unless there is some strong reason for expecting a change. We may suppose, then, that the first approximation to a best estimate of average GNP over, say, ten years is that it will be the same as it is now. That forecast will then be "corrected" in various ways, depending on the recent history of movements of income and on the position of other variables which are commonly supposed to influence GNP movements.

The forecast will tend to be corrected downward if the current level of GNP has been reached in a sudden burst of growth, particularly if this was due to obviously temporary factors, such as a short-term movement in government expenditure.

On the other hand, there may be an upward correction for trend, particularly if income has been moving upward fairly steadily for some years. An erratic upward movement will lead to a much weaker trend projection.

If there has been a deep depression in the fairly recent past, some firms may consider a repetition within the next few years to be very likely, and for those firms the best forecast will be corrected downward. In most cases, however, fear of a depression will enter expectations by creating uncertainty rather than by changing the modal point of the distribution.

If income has already declined from a previously achieved level, many firms will fear a major depression, and best forecasts will be below current levels of demand. Probability distributions may have a wide variance and be fairly symmetrical because firms will be torn between projecting the downward trend and predicting a return to normal.

[8] E. Cary Brown, "Some Evidence in Business Expectations," *Review of Economics and Statistics*, vol. 31, 1949.

[9] Charles F. Roos, "Survey of Economic Forecasting Techniques," *Econometrica*, vol. 23, no. 4, October, 1955.

Without taking anything into account except the past history of income, it appears reasonable to suppose that the mode of a decision maker's probability distribution of average future incomes will tend in most cases to be fairly near the present level. The possibility of deep depression on the one hand and of upward trend on the other will determine the shape of the distribution. A long period of rising income interrupted only by minor depressions will skew the distributions to the right by giving a heavy weight to trend projection and a light weight to fear of a depression. The skewness will go in the other direction in periods immediately after a major depression.

If a decline in income is already under way, the variance of the probability distribution will become large as fears of a depression increase the weight of negative departures from the present level, while hopes of recovery add to trend experience in increasing the weight on the other side.

Expectations about year-to-year variations in demand are presumably conditioned chiefly by the experience of the past few years. They will also be influenced by the degree of faith which the business community places in the adequacy of the stabilization measures available to the government and the willingness of the authorities to apply them.

Expectations about the possibility of a depression in the near future are conditioned by two types of consideration. Many forecasting efforts are strongly influenced by the signs of the times, e.g., movements of profits, numbers of business failures, and unemployment. All of those variables tend to move with the rate of increase of income. In addition, rates of increase of sales have an important direct influence on expectations. Thus short-term expectations should be influenced by movements in the rate of change of income.

Short-term expectations will also be influenced by a variety of more or less random factors—political events, random movements in the financial markets, and the rise and fall of speculative activity in particular industries.

The expectations of individual firms or groups of firms about the prospective movement of demand for their products relative to the movement of aggregate demand will be very strongly influenced by the observed past rates of growth of demand for their products. That influence will, of course, be modified by considerations which lead the firms to believe that the past growth was in some way atypical and by some direct information on future demand prospects, e.g., market analysis of uses for a product or "saturation" analyses.

D. Effect of Income Movements on Marginal Efficiency

In section B we discussed the effects of various patterns of expectations on the marginal efficiency of capital, and in the last section we considered the effect of income movements on the pattern of expectations. The two sections can now be linked to obtain the relations between past income movements and current investment.

It has been argued that there is a strong tendency to project the current level of demand into the future so that the marginal efficiency of capital is primarily determined by the current level of demand (and, of course, capital stock and the other factors mentioned above), but the marginal efficiency based on the projection of current levels of demand is modified in various ways by the influence of past movements of demand.

If the income has had a high average rate of growth in the past few years, firms will tend to expect that the average of future incomes will exceed the current value of income. For firms using indivisible equipment, the marginal efficiency of a given amount of investment will in those circumstances tend to exceed the marginal efficiency obtained from assuming a constant level of demand. The effect of a high rate of growth will be stronger if growth has been steady than if it has been erratic. The upward shift in the marginal-efficiency schedule which results from a positive past rate of growth is probably fairly small. The average rate of growth of GNP (in real terms) seldom exceeds 4 per cent for more than a couple of years, and the growth is always erratic. The tendency to build new indivisible plants on the basis of growth projection (of GNP as distinguished from growth of demand for particular products) is likely to be somewhat weak because the growth is slow and uncertain.

A high past rate of growth will also tend to raise the marginal-efficiency-of-investment schedule because the danger of being caught with excess capacity in a depression is smaller when the trend of demand is upward.

A fairly deep depression in the recent past tends to lower the marginal efficiency of investment by skewing the distribution of expected future incomes in the downward direction. The occurrence of depressions of any type in the record tends to reduce marginal efficiency by increasing the need for flexibility and increasing the expected effect of depreciation charges on reported net income.

Changes in the rate of growth in the immediate past influence marginal efficiency through their influence on near-term expectations.

Obviously a downward movement of income in the recent past tends to reduce marginal efficiency of investment.

III. INVESTMENT IN OTHER LINES

So far we have been assuming that all the investment made by the firm is aimed at the production of some fixed line of products. In that case the shape and level of the firm's marginal-efficiency-of-investment schedule is determined by the level of demand for its products, the technical properties of its existing capital stock, the technical properties of capital goods which it could buy, and the prices of capital goods and other input factors. It is useful to start in that way, but it is important to recognize that firms are free to invest in facilities for producing products which are new to the firm. They do so by buying control of existing firms in other lines, by building new plant for the production of additional products, or by buying small existing firms in other lines and then expanding their facilities. Thus investment in new product lines by an existing firm may or may not contribute to the volume of real investment and to the generation of income. Of course, even when one firm uses its funds to buy other firms, it may influence total investment and expenditure, since the recipients may use the funds for real investment or buy securities or spend them on consumption. In that case, however, the income effects of the investments are much less certain than in the case of real investments.

It is not only true that firms are free to invest in lines other than the line in which they are engaged, but there are a number of positive reasons why they should wish to do so. These reasons are of two sorts: (1) those connected with the institutional arrangements for corporate finance and (2) those connected with the competitive position of firms.

There is bound to be some tendency for the marginal cost of raising funds, and therefore the marginal efficiency of capital, to be higher in rapidly growing industries than in slowly growing ones.[10] There is some inducement, therefore, for firms in slowly growing industries (provided that the competitive structure permits good profits in spite of slow growth) to move into more rapidly growing sectors, the marginal efficiency of capital being higher in the high-growth sector than the marginal cost of capital in the low-growth sector.

That consideration is reinforced in the case of relatively small firms in a growing industry. The marginal cost of capital to them is pushed up not only by their growth but also by their inability to issue securities on the organized exchanges. The prospective value of the assets and market position of a growing small firm is greater to a large firm in a slowly growing industry which can raise capital easily than it is to the owners.[11]

If the owners of a small firm have to raise cash to pay estate taxes or wish to convert a prospective income stream into a capital gain for tax purposes, the gap between the prospective value of the small growing firm to its owners and its value to a large firm in a slowly growing industry is widened still further.[12]

The active role is not always played by the slowly growing firms. Firms in need of cash have sometimes arranged mergers with those

[10] Firms in rapidly growing industries can obtain outside capital more cheaply than firms in slowly growing industries. However, firms in slowly growing industries usually do not need outside funds, while those in rapidly growing industries do. Internal funds almost always have a lower imputed cost than outside funds.

[11] J. K. Butters and J. V. Lintner, *Effects of Taxation in Corporate Mergers*, Division of Research, Graduate School of Business Administration, Harvard University, Boston, 1951.

[12] *Ibid.*

which were oversupplied with cash. In recent years there have been a number of mergers for tax purposes between firms with profits and firms with tax-deductible losses.

The second group of special motivations for branching out is connected with competitive strategy. A firm may move into new lines by integrating forward or backward by producing additional products which make use of the same raw materials, production equipment, or distribution facilities as those already being produced.

Firms in basic industries find it advantageous to integrate forward because by doing so they ensure a more stable demand for their product. Instead of competing with other producers in their own industry for sales to fabricators, they compete with fabricators for sales of the final product. To the extent that customer attachments are stronger in consumers'-goods lines than in raw materials, they have a more stable demand when they fabricate. Of course, if they do not fabricate their whole output, they may lose the goodwill of fabricator customers by competing with them. Basic-materials firms may also go into fabrication in order to push new uses for the material they produce. Finally, when the materials producers are better financed than the fabricators, the financial considerations mentioned earlier apply.

Firms that use competitively produced raw materials may find it desirable to integrate backward in order to reduce the risk of fluctuations in the price of their raw materials. In other cases, firms may attempt to acquire raw-material sources required by competitors in order to improve their bargaining power, or they may acquire their own sources to prevent others from doing so.

Firms that produce an "assembled" final product may integrate backward to ensure the quality and regularity of supply of components. Or, when they are larger and more efficient than their own suppliers, they may integrate backward to reduce costs.

Horizontal integrations are specially advantageous when the new product lines use similar raw materials or similar processes or are sold through the same distribution channels. Firms that own raw-material resources for the production of one group of products may be able to exploit those resources more fully by taking on more products which use them. The rationale is the same as that of forward integration discussed above.

When demand for a firm's products is seasonally variable or erratic, it can get fuller utilization of its facilities by widening its line so as to produce products with a different seasonal pattern or to average out the random fluctuations in demand provided that the new products utilize at least some of the equipment required for the old ones.

Adding new products which use the same distribution facilities as those already produced has a number of advantages: (1) There are economics of scale in the use of salesmen, trucks, and warehouses, and possibly in accounting and advertising. (2) The goodwill of the firm can be exploited more fully. (3) Production of a full line of products makes it possible to enter into exclusive-dealer arrangements and freeze competitors out of certain outlets. It also makes the line more attractive to dealers.

Firms may be led into horizontal integration because their research efforts have produced new products or processes which can be exploited more profitably directly than by sale or licensing of patents.

Two other reasons for integration are of a different order. Firms in cyclically unstable lines may wish to reduce fluctuations in income by producing products with a steady demand. They may consider this worthwhile even if there are no other ancillary advantages and even if the prospective average returns in the new line are no higher than in the old.

Finally, when high profits are earned by a monopoly or a well-organized oligopoly, other firms may try to break into the charmed circle unless the monopolists have held down their profits in order to avoid that possibility.

Those different motivations for integration do not operate independently but reinforce one another, and in any particular case two or three of them will be operating together.

In general, it can be said that the rationale of integration is based on one or both of two considerations. (1) Monopoly or differences in the balance of supply and demand for funds may make the marginal efficiency of capital higher in one industry than in another and therefore encourage a flow of funds from one field to another. (2) For a variety of reasons it is possible that the marginal efficiency of investment in a given industry may be higher to a firm in another industry than to the firms already in the industry in question. Indeed this can be a symmetrical situation. The marginal efficiency of investment in industry *A* may be 10 per cent to firms in *A* and 15 per cent to those in *B*, while the marginal efficiency of investment in *B* may be 10 per cent to those in *B* and 15 per cent to those in *A*.

The financial aspects of initial investments in new lines of products are rather different from those of investments in the firm's established lines. Many mergers are accomplished by an exchange of securities without requiring any cash outlay on the part of the acquiring company. The occurrence of such mergers affects the movement of investment at later dates, but they are not competitive with other investment at the time at which they occur.

Other mergers may involve a cash outlay on the part of the acquiring firm but may or may not be competitive with other uses for the purchaser's funds or borrowing power. If a firm issues securities or if it uses its cash to buy another firm which has more stable earnings than its own or which promises to produce a very high yield on the purchase price, the purchaser's borrowing power (with a given degree of risk) may be increased by more than the amount laid out to accomplish the merger. In that case the acquisition of the other firm, far from competing with real investment, encourages the firm to increase its real investment.

In general, the amount of funds absorbed by mergers is probably relatively small. Only small companies will be purchased for cash. Though the number of such purchases is sometimes fairly large, the amount of money involved is relatively small.

Mergers effected by exchange of securities are interesting because of their effect on competitive structure, but they have no direct effect on investment.

Real investment in products new to the firm, whether new to the economy or not, is the important category. The existence of ancillary advantages from integration or diversification implies that the aggregate marginal-efficiency-of-investment schedule will be at a higher level for a given distribution of real capital among firms and a given level of aggregate income than it would be if each firm had to stay in its own territory. For the economy as a whole the level of returns to be had from real investment in new product lines will vary in the same way as the returns from investing in existing lines, i.e., with aggregate demand and capital stocks.

However, there will be some differences. Severe competition may make integration of some types more advantageous in a depression than in a boom, but it can then be accomplished by purchase of other firms. On the other hand, the need to strengthen competitive position is least when demand is very high relative to capital stock. Moreover, the returns from integration and diversification are likely to be long-term ones, because it takes time to develop engineering and managerial technique and to develop a market position. Many firms will therefore be unwilling to borrow to move into new lines even though they feel fairly certain of making a good return from such investments in the long run.

As a result, firms may postpone some investment in new lines until they can be financed from internal sources. Consequently the greatest amount of investment of the type in question may occur in years of high prosperity but not in years in which income is growing at a very rapid rate.

The fact that firms can integrate or diversify influences the rate of investment in a number of ways. (1) Because of the "side effects" of integration and diversification the amount of real investment which will yield a given prospective return is increased for most firms. (2) The acquisition of other firms may compete for funds with real investment. In that case the favorable influences of integration and diversification on rcal investment are partially offset. However, it seems likely that competition for funds between mergers and real investment is not very important. (3) The elasticity of the marginal-efficiency-of-investment schedule for individual firms should be much greater when the schedule takes account of opportunities for diversification than when it does not. Normally, the marginal efficiency of investment (in a given industry) will be below the average rate of return on capital in that industry. It will fall as the rate of investment increases. But one firm can enter another industry and invest a large amount of capital at a prospective rate of return only a little below the average rate of return in the industry being entered. As a result the marginal efficiency of investment in diversification should fall very slowly with the rate of investment.

IV. SUMMARY AND CONCLUSIONS

The argument of this chapter leads to the following conclusions: (1) Industrial firms are faced with schedules of marginal efficiency of investment which slope downward gradually as the rate of investment is increased. (2) The schedule moves upward and outward as demand increases relative to capital stock. It will also be influenced by expectational factors, some of which are themselves influenced by past movements of income. Prices of capital goods and other inputs and rates of technical change will also be influential. Finally, opportunities for diversification and integration tend to raise the amount of investment which will yield a given return.

If the firm wishes to maximize its profits, it ought to invest an amount which will just equate the marginal cost of raising capital to the marginal efficiency of investment.

In the next chapter we shall consider the factors which influence the marginal cost of raising capital. We will then be in a position to consider the interaction between the factors influencing the yield from investment and those influencing the cost of capital.

CHAPTER 5

Sources of Funds and the Cost of Capital

In the last chapter it was shown that the schedule of marginal efficiency of investment for the individual firm slopes downward and has an appreciable elasticity (over the range of rates of investment which is likely to be relevant). A firm will maximize profits if it invests at rates which equate the marginal efficiency of investment to a discount rate which represents the marginal cost of raising capital. Until fairly recently it was customary to identify the marginal cost of raising capital to a firm with the market rate of interest which the firm must pay. Most large firms could borrow a large amount (in any one year) relative to a single year's rate of investment without forcing up the interest rate they would have to pay. That seems to justify us in treating the supply of capital to the firm as perfectly elastic, but doing that leads to results which are not easy to reconcile with the facts.

If the supply of capital to the firm were perfectly elastic, investment would be on a hand-to-mouth basis. Investment would vary with the factors which shift the marginal-efficiency schedule upward and to the right or downward and to the left. Other things equal, a large amount of net investment would occur in periods when income rose rapidly, and small amounts in periods when income rose slowly. That is the justification of the acceleration principle. There would be nothing to prevent firms from exploiting investment opportunities as quickly as they appear (except limitations on the supply of physical capital goods or engineering staffs).

In fact investment moves as though it were being backlogged in the years in which investment opportunities are very large and released in later years. The period 1946–1949 is an example. During that period the ratio of gross stock of producers' durables to output rose gradually from about .52 to .66.[1] At the same time the average age of the equipment fell. That change was not, of course, due to technical change. It was due to the increase in income during the war, when capital formation was restricted. The investment opportunities generated by the increase in income during the war were exploited gradually for a number of reasons. Investment was restricted in some fields by the supply of capital goods. The investment programs of some firms were restricted by lack of engineering staff. Some firms restricted their investment for a time because they feared a postwar depression, but, in addition, some firms restricted their investment because they lacked the necessary funds.

The postwar situation is a special one, but there is other evidence of backlogging. During the recession of 1953–1954, investment continued at high levels. Investment in producers' durable equipment fell from $25.0 billions (current prices) in the third quarter of 1953 to $21.9 billions in the fourth quarter of 1954, or about 12 per cent.[2]

Expectations were worse during the recession than previously. We cannot, therefore, attribute the continuation of investment to an improvement in expectations. There were only minor shortages of capital goods in the period just preceding the recession. Finally, we cannot attribute the high level of investment to the completion of projects undertaken in earlier years. Construction contracts awarded for manufacturing in the first half of 1954 were only 22 per cent below the awards of the first six months of 1953. In the utilities field, contract awards in the first half of 1954 were 10 per cent above those of the first half of 1953. New orders for durables fell by about 23 per cent from the first half of 1953 to the first half of 1954.[3] There is a strong case, therefore, for supposing that investment opportunities which could have been exploited before the recession were being exploited during the recession. It seems reasonable to suppose

[1] R. Nassenbein and D. G. Wood, "Growth of Business Capital Equipment," *Survey of Current Business*, December, 1954.

[2] *Economic Report of the President*, January, 1957.

[3] Construction contract data from *Federal Reserve Bulletin*, January, 1955, p. 71, and January, 1954, p. 83. New-order data from *Economic Report of the President*, January, 1955.

that limitations on the supply of money capital were responsible for the delay.

It is always dangerous to suppose that one explanation of a fact must be correct because no other is available. In this case, however, there is other evidence of the role of financial factors in determining the rate of investment. Meyer and Kuh have recently published a study of the investment behavior of a large cross section of firms.[4] The study covers the years 1946–1950. Over 750 firms, divided into ten industry groups, were studied. Multiple-correlation methods were used to explain the variation among the rates of investment of the firms in each industry group in each of five years. The results show that variables measuring the availability of funds for investment—profits, liquid assets, depreciation expense—play an important role in the explanation in all industries and years. The influence of the "liquidity variables" is compared with a measure of the intensity of utilization of capital stock. The authors conclude that the liquidity variables were relatively more important in the later years than in 1946 and 1947. The explanation given is that liquid assets and profits were relatively higher in the early years. A summary of the results is given in Table 1.

The Meyer and Kuh study gives fairly convincing evidence of the role played by financial variables. Klein and Goldberger have obtained similar results on a time-series basis.[5] Tinbergen found that profits played an important role in the explanation of investment in his 1919–1932 study of investment in the United States.[6] Klein found that profits were important in the explanation of railroad investment (in a cross-section study).[7]

Because the factors determining profits are closely related to the

[4] J. Meyer and E. Kuh, "Acceleration and Related Theories of Investment," *Review of Economics and Statistics*, vol. 37, 1955. See also J. Meyer and E. Kuh, *The Investment Decision*, Harvard University Press, Cambridge, 1957.

[5] L. R. Klein and A. S. Goldberger, *An Econometric Model of the United States 1929–1952*, North Holland Publishing Co., Amsterdam, 1955.

[6] Jan Tinbergen, *Business Cycles in the United States 1919–1932*, League of Nations, Geneva, 1939.

[7] L. R. Klein, "Studies in Investment Behavior," in *Conference on Business Cycles*, National Bureau of Economic Research, New York, 1951.

factors determining the marginal efficiency of investment we cannot suppose that profits are the sole determinant of the rate of investment. But the studies cited are certainly consistent with the view that financial factors play an important role in determining the rate of investment.

That view is inconsistent with the idea that the supply of capital to the firm is very elastic. Large firms, at least, can certainly increase their debt by a substantial amount without forcing up the interest rate they have to pay. Some explanation is therefore required. It is argued below that the imputed cost of raising funds rises much more rapidly than the market rate of interest. It is also argued that the marginal-efficiency-of-investment schedule for most manufacturing firms intersects the supply-of-funds schedule in the inelastic range. The position of the inelastic range of the supply-of-funds schedule is largely determined by the level of profits (in dollars per year). It follows that the rate of investment will be strongly influenced by changes in profits.

We proceed in the following way. In section I the cash and imputed costs of raising funds by various means are considered. In section II we derive a marginal-cost-of-funds schedule, showing that the marginal cost of raising funds depends on the amount raised, the flow of retained earnings, the amount of debt outstanding, and the terms on which equities can be issued. The factors causing differences in the shapes of the schedules in different industries are outlined. In section III the possible positions of the intersection of the marginal-efficiency-of-investment and marginal-cost-of-funds schedules are indicated. Up to that point the amount of profits and retained earnings available to the firm have been taken as given. In section IV it is shown that there is a strong connection between dividends, profits, and lagged dividends. In section V we consider the theory of profits. The theory shows why it is likely that most manufacturing firms will operate in the inelastic range of their cost-of-funds schedules. In section VI the argument of Chapters 4 and 5 is summarized and put into the form of a theory of business investment. In section VII we draw some implications for business-cycle theory.

I. SOURCE OF FUNDS

The supply of funds for investment available to any firm comes from several different sources: (*a*) depreciation allowances, (*b*) retained net earnings, (*c*) various types of borrowing, (*d*) sales of equities. Each of these sources of funds has a cost in some sense, but the nature of the cost varies with the source and with the expectations of the management and the lenders. Even under conditions of certainty, debt sources have a cost in the form of interest. The investment of depreciation allowances and retained earnings does not involve the firm in any cash costs, but it does involve an opportunity cost. The firm could always lend its depreciation allowances and retained earnings. In addition, the retention of net earnings involves the firm in imputed costs of two other types. If the management is concerned about the value of the stockholders' equity, a cost must be imputed to the retention of earnings, because a dollar of retained earnings does not have the same effect on the value of the stock as a dollar of dividends.[8] Even if the management is not concerned about the value of the stockholders' equity, it must be concerned about the complaints of stockholders who may want more dividends and about the risk of losing control as a result of stockholder dissatisfaction. The imputed cost of funds from retained earnings must be supposed to rise as the proportion of earnings retained rises.

The sale of equities also involves imputed costs. An equity issue will reduce earnings per share unless the proceeds can be invested at a yield equal to or greater than the ratio of average expected future earnings to the issue price. Managements are generally concerned about their earnings record on a per share basis, and will be unwilling to finance investments by equity issues unless they can expect an improvement in the per share earnings.[9]

[8] J. V. Lintner, "Distribution of Incomes of Corporations among Dividends, Retained Earnings and Taxes," *American Economic Review*, May, 1956 (*Papers and Proceedings of the American Economic Association*, December, 1955).

[9] David Durand, "Costs of Debt and Equity Funds for Business," in *Conference on Research in Business Finance*, National Bureau of Economic Research, New York, 1952.

The use of borrowed funds has a cash cost in terms of interest payments, but it generates additional imputed costs. Fixed charges increase the effect of fluctuations in demand on reported earnings and on the amount of cash available for dividends. Fixed charges and repayment commitments may lead to a situation in which the firm's cash inflows are insufficient to meet commitments to pay cash. The firm must then become a necessitous borrower and accept terms and conditions from lenders which restrict the management's freedom of activity. In the extreme case it may go bankrupt.

TABLE 1

ANNUAL AVERAGES OF PARTIAL CORRELATIONS BETWEEN INVESTMENT AND OTHER VARIABLES

Year	Variables					
	Profits	Depreciation expense	Sales	Net quick liquidity	Change in sales	Capacity
1946	.007	.083	.076	.012	.0296	.452
1947	.073	.066	.056	.130	−.0624	.485
1948	.217	.093	.206	−.043	−.0235	.145
1949	.266	.259	.255	−.084	.0021	.016
1950	.210	.280	.181	.003	.0076	.042

Source: J. Meyer and E. Kuh, "Acceleration and Related Theories of Investment," *Review of Economics and Statistics*, vol. 37, 1955.

Managements are obviously averse to becoming necessitous borrowers or going bankrupt. When a firm has debt outstanding, it will not be willing to make an investment whose probable yield is insufficient to cover a risk premium for the increase in the probability of getting into financial difficulty generated by the increase in fixed charges associated with the investment (or the failure to reduce fixed charges when the investment is financed from earnings). Even when the risk of getting into financial difficulties is negligible, the yield of an investment must be sufficient to cover the imputed cost of an increase in the range of fluctuation of reported earnings and amount of funds available for dividends.

The risks associated with borrowing will also be reflected in the value of the equity. If the management is interested in the stockholders' welfare, or if it contemplates selling equities, the same set of risks may be reflected again in the imputed costs of debt finance.

It seems clear that the yields required to justify investments financed with additional debt or with funds which could be used to pay off debt will rise with the ratio of the interest charge on existing debt to earnings.

II. THE COST OF FUNDS FOR INVESTMENT

With those general considerations in mind let us attempt to construct a schedule of the marginal cost of funds for investment. The problem is complicated at the outset by the presence of a stock-flow interaction problem. Retained earnings are a flow, while debt is a stock which may be increased or decreased by borrowing or repayment. As long as it has debt, the firm is paying interest and taking risks, and the imputed cost varies with the amount of debt outstanding. We can relate the flow of retained earnings to the stock of debt by considering that the cost of using retained earnings for real investment is the cost of not repaying debt.

In this section we shall first consider the opportunity cost of using retained earnings for real investment on the assumption that the firm has no debt. We then consider the relation between the opportunity of using retained earnings and the existing stock of debt. We then consider the cost of issuing equities. Finally, we consider the short-run marginal-cost-of-funds schedule for a firm with a given amount of debt.

A. Opportunity Cost of Using Retained Earnings (No Debt)

If we suppose for the moment that dividend payments are independent of capital requirements, every firm will have available in each year a block of funds equal to the sum of depreciation expense and retained net earnings from the previous year. The cost of utilizing those funds for real investment is an opportunity cost—the return which the firm could obtain by using the funds to repay debt or to buy the shares or assets of other companies. As we have already seen, the returns from buying other firms with certain strategic advantages may be relatively high, but the supply of such opportunities is limited. The return from other equities is the dividend yield plus a problematical capital gain from future increases in dividends generated by returns on retained earnings. The dividend return is presumably as safe as or even somewhat safer than the return from direct investment (if the companies whose shares are purchased are similar to the one in question). But there is a market risk involved, at least for bookkeeping purposes. In addition, most managements are likely to feel that they would rather worry about their own business than about someone else's. When those considerations are taken into account, the opportunity cost of internal funds cannot be much above, and may well be below, the dividend yield on listed equities, which is usually of the order of 5 or 6 per cent. Indeed, since relatively few firms ever hold marketable equities (without obtaining some measure of control), it must be supposed that the prejudice against playing the stock market is sufficiently strong to put the valuation of returns on equities down to the level of the returns on bonds.

In other words, the opportunity cost of using retained earnings for real investment is only slightly above the yield on bonds, provided the firm has only a small amount of debt relative to earnings.

B. Cost of Debt

So far we have been neglecting the existence of debt. But if a firm already has debt, the opportunity cost of using retained earnings will be above the yield obtainable by buying securities, for, if it repays debt or holds liquid assets to balance it, there is a saving in interest charges plus a reduction in risk.

The return from repaying debt or holding liquid assets to balance it will be equal to the yield on bonds as long as the ratio of debt service to earnings is fairly low. As debt rises relative to earnings the risk premium required to cover the leverage of debt service on earnings fluctuations will increase. The opportunity cost of not repaying debt will therefore increase. That opportunity cost increases the return required to justify investing in physical assets whether the investment involves taking on additional debt or failing to repay existing debt.

The supply curve for investable funds in a single year can be obtained in the following way:

Let the imputed marginal cost of debt be $f\left(\frac{D}{\pi}\right)$, where D is the amount of debt and π the rate of profits (in dollars per year). If the firm invests nothing in the year, then its debt at the end of the year will be $D_t - RE_t - A_t$, where D_t is the initial debt, RE_t is anticipated retained earnings, and A_t is depreciation expense. The opportunity cost of the first dollar of investment for the year is $f\left(\frac{D_t - RE_t - A_t}{\pi_t}\right)$, and in general the opportunity cost of the last out of I dollars of investment is $f\left(\frac{D_t - RE_t - A_t + I_t}{\pi_t}\right)$. This holds provided $D_t - RE_t - A_t + I_t$ is positive. If that quantity is negative, the supply price of investable funds is of the order of the yield on bonds for the reasons given above. For purposes of this argument the profit figure should be an average of past profits rather than the current profits.

Since firms are reluctant to increase the variance of net earnings, the cost of funds will begin to rise sharply as the ratio of debt to earnings rises.

C. Equity Issues

The imputed cost of issuing equities is substantially above the opportunity cost of using internally generated funds or small amounts of debt. If current earnings can be regarded as bearing any relation to the earnings expected in the future by the management, the yield cost of equity finance is usually of the order of 7 to 10 per cent for large firms. To this must be added flotation costs plus any reduction in the value of existing shares resulting from the issue. The differential is further increased by the differential tax treatment of bond and equity finance.[10]

The cost-of-funds schedule will be the one discussed above until

[10] Durand, *op. cit.*

the volume of outstanding debt becomes great enough to push the imputed cost of debt finance up to the level required to cover the cost of issuing equities. Occasionally a stock-market boom may create a situation in which equity yields are very little higher than bond yields, so that the imputed cost of bond finance is above the cost of equity finance. For the most part, however, equities will be issued by firms whose growth rate is high enough relative to profits to induce very heavy use of outside funds. Public utilities are the best example of that class of firm. Most large manufacturing firms are not in that category, though many small and moderate-sized firms are.

Once returns are high enough to justify issuing equities, the supply-of-funds schedule becomes more elastic to the rate of return than in the range in which there is a large volume of debt outstanding but equities are not being sold. However, the supply schedule for equity funds is by no means perfectly elastic, because large new issues tend to depress the price of the security.

D. Short-run Cost-of-funds Schedule

If we plot the imputed cost of investable funds against the rate of investment with fixed retained earnings and profits, we get a schedule like the one in Figure 8, where the imputed cost of funds is measured on the vertical axis and the rate of investment plus initial debt on the horizontal axis. The schedule is perfectly elastic, or nearly so, at a relatively low rate of return until $I + D$ equals depreciation expense plus retained earnings. It then rises, at first slowly and then steeply as investment plus initial debt increases. When imputed cost of debt rises so high that equity issues are justified, the slope is reduced again. To obtain a schedule of investable funds for a particular year, we move the origin to the right by an

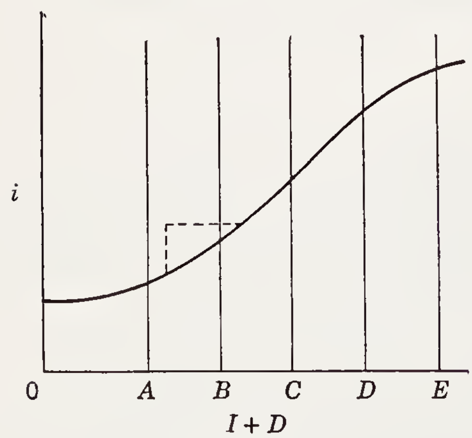

FIGURE 8.

amount equal to the firm's initial debt and consider only a limited range of the schedule.

The curve in Figure 8 was drawn on the assumption of fixed retained earnings and expected profits. An increase in retained earnings shifts the whole curve to the right by an equal amount. An increase in the level of expected profits also shifts the curve to the right. An increase in profits which persists first increases retained earnings by a large amount because of the dividend lag, but has little effect on expected earnings over longer periods. As time passes, the retained earnings fall as dividends catch up, while the security of the gain in profits has a favorable effect on expected earnings.

During a period of income growth the whole supply schedule slowly shifts to the right. At the same time a firm which increases its real assets in proportion to demand will have a marginal-efficiency schedule which stays at about the same level for zero investment and slowly shifts rightward (or flattens out) as the scale of the firm grows. In that case the firm will continue to operate within a segment of the cost-of-funds curve having the same elasticity.

Of course, since demand actually grows erratically, the intersection of marginal-efficiency schedule and cost-of-funds schedule will rise and fall. But firms whose rate of growth relative to the growth of income is roughly constant will continue to have the same type of short-run supply-of-funds schedule as long as the average (over three or four years) rate of growth of income is steady.

E. Industry Differences

Of course, the shape of the curve is not the same for all firms. The width of the elastic range will vary with the efficiency and competitive position of the firm. It will also vary with the dividend policy of the firm. Similarly, the rate at which the imputed cost of debt rises with debt will vary with the nature of the business and the attitudes of management toward risk. Some firms have a rooted prejudice against debt. The imputed-cost curve will therefore rise very rapidly for those firms until it reaches the level at which equities can be issued. Firms which do not have a rigid attitude about debt will have varying degrees of reluctance to take on debt. This reluctance is often expressed in terms of a "safe" ratio of debt to earnings or of debt to equity capital. That means that so long as the debt ratio is below a certain figure the imputed cost of debt is not much above the market rate of interest. As the ratio nears the critical figure the marginal imputed cost of debt rises sharply. The imputed cost of debt will vary with the nature of the business as well as with the attitudes of managements. In businesses which are subject to a great deal of cyclical risk or to rapid technological change, debt finance will be avoided. Businesses which are safe both competitively and with respect to cyclical fluctuations can safely carry a good deal of debt. Public utilities are in the latter range.

At the time of the 1956 steel strike the large steel companies maintained that an increase in profit margins was necessary to finance expansion. Rationalization aside, they were, in effect, arguing that the steel business is too risky to permit debt finance. That is an argument which cannot really be tested, since risk, like beauty, is in the eye of the beholder.

The terms on which equity capital can be raised will vary with the riskiness of the business and with its growth prospects. Again, public utilities are able to raise equity capital on more favorable terms than are most other businesses.

The cost-of-funds curves for different firms will therefore vary widely in level and shape. The cost-of-funds curve for a utility company is likely to be fairly elastic over a wide range, and the imputed cost of funds may be relatively low even when the ratio of investment to retained earnings is quite high.

A conservative steel firm, on the other hand, will have a cost-of-funds curve which slopes steeply upward as soon as investment exceeds retained earnings and depreciation by a relatively small amount. Such a firm tends either to rely entirely on internal funds and a relatively small amount of debt or to issue equities. It will issue equities, however, only when the marginal efficiency of investment is quite high.

For such firms (and they may be a majority among manufacturing firms) the supply-of-funds schedule has a very simple form. The schedule will be elastic at a low level of returns from zero gross investment up to a certain point. That point can be found by taking the sum of (*a*) depreciation expense, (*b*) retained earnings, and (*c*) the difference between a safe amount of debt and the existing amount. After that the schedule rises sharply to the yield which justifies the issuance of equities. Item *c* above may, of course, be either positive or negative, according to whether profits have been increasing or decreasing. But in prosperous times, when profits are rising, it will usually be positive.

There is, of course, nothing sacred about debt limits, and management attitudes about them are likely to change from time to time as their faith in future prosperity waxes and wanes. A company may, of course, change its views if it requires funds to expand enough to keep its market share. However, in those circumstances the yield on investment is likely to become high enough to justify the sale of equities.

III. COST OF FUNDS AND RATE OF INVESTMENT

With all those considerations in mind we are now in a position to work out the interaction between the marginal-cost-of-funds schedule and the marginal-efficiency-of-investment schedule. Four types of position for the intersection can occur.

1. For firms like public utilities which can issue equities relatively cheaply, are growing rapidly, and have relatively small retained earnings, the marginal-efficiency schedule (without taking account of diversification) may usually intersect the marginal-cost-of-funds schedule in the range in which it pays to issue equities.

2. Fairly rapidly growing firms may find that the marginal-efficiency schedule (without taking account of diversification) intersects the marginal-cost-of-funds schedule in the inelastic range.

3. Other firms may find that the intersection occurs in the inelastic range only when the marginal-efficiency schedule is shifted upward by taking account of the possibilities of diversification. Some firms in this class will borrow up to their debt limit. Others,

however, will engage in diversification only to the extent that they can do so without borrowing.

4. Finally, some firms may not be able to invest their retained earnings at more than an opportunity-cost yield even if they do diversify. Since in most cases very large amounts can be invested in other industries, the firms in this latter category will be few in number. They will include large, extremely profitable firms whose managerial resources will not permit further extension or those who fear political repercussions from further expansion.

It follows from the argument above that most firms will be operating just at their debt limits when income is rising. A few will be issuing equities, and a few will have no debt. In the latter two cases the rate of investment will depend primarily on the factors determining the rate of investment at which marginal efficiency equals cost of funds from equities in the one case and opportunity cost of retaining earnings in the other. But for most firms the rate of investment will depend on the level of profits, the allowable debt ratio, and the proportion of earnings retained. When profits, rise the firm will be able to invest its retained earnings plus an increase in debt. When profits fall, the firm will be able to invest retained earnings minus a decrease in debt (or increase in liquid assets).

IV. DETERMINATION OF RETAINED EARNINGS

So far we have simply assumed that the amount of retained earnings is fixed. We must now consider the determination of retained earnings out of given profits. We can then consider the determinants of profits themselves. One would expect that dividend policy would be related to the opportunity to invest retained earnings. No doubt there is some connection, but it appears to be a relatively weak one.

Professor Lintner's[11] studies of dividend policies indicate that most firms have some idea of a normal ratio of dividends to earnings. When earnings change, dividends are adjusted upward or downward toward the amount normal for the new rate of earnings. The

[11] *Op. cit.*

adjustment is a gradual one, however, so that dividend payments depend on past dividends as well as on profits.

Lintner explains dividends by the use of an equation of the form $d_t = \pi_{t-1} + bd_{t-1}$ (d_t = dividends in period t, π_t = profits in period t). He argues that most manufacturing firms follow a dividend policy which is independent of their capital requirements.

There are two reasons why year-to-year variations in capital requirements may not influence dividend policy: (1) It is easy to understand that firms are conservative about increasing dividends in response to a rapid increase in profits. If dividends are allowed to vary with profits, it will often be necessary to reduce them. Most firms would rather raise dividends slowly than run the risk of first raising them and then lowering them. (2) If dividends lag behind earnings, retained earnings will increase rapidly when profits increase. Profits are correlated with investment opportunities, so that a passive policy generates movements of the supply of retained earnings which are correlated with the demand for investable funds. Firms may need still more funds in a boom, but they are naturally reluctant to reduce dividends at the very moment when profits are increasing.

As between firms, the rapidly growing firms which have the greatest need for funds tend to have the highest profit rates. That helps to supply the funds they need. In addition, the earnings of a rapidly growing firm will increase rapidly. If such a firm follows the same dividend-adjustment formula as a more slowly growing firm, it will nonetheless pay out a lower proportion of its earnings because the dividend-adjustment lag will have a greater effect.

On the strength of Professor Lintner's very thorough study we shall continue to assume that dividends are determined independently of prospective returns on the investment of retained earnings.

V. PROFITS

The determination of profits is a much more complicated matter. No very satisfactory theory of corporate profits exists, so it is necessary to devise one. The problem divides naturally into two parts. First we must consider the relation between profits and the utilization of capacity. Then we can consider the interaction between profits, investment policy, and rate of growth of demand.

A. Capacity Utilization and Profits

We usually expect that the ratio of both net and gross profits to invested capital will decline as capacity utilization declines. Gross profits per unit of capital will fall with capacity utilization because markups over direct cost fall with utilization. Net profits will fall even faster because of the influence of depreciation and interest changes. The elasticity of profit rates with respect to utilization rates will vary with competitive conditions. In purely competitive industries, markups tend to fall rapidly with utilization, while they may fall very little for monopolies or well-organized oligopolies. In most industries (other than purely competitive ones) markups usually do not increase beyond a certain limit even when current demand cannot be supplied.

We can use those very elementary observations as a basis for a discussion of the mutual adjustments between investment, growth of demand, and profits. We can assume that whenever the firms in an industry increase capacity at a rate faster than the rate of growth of demand, the rate of return on capital will fall, and conversely.

B. Profits and Rates of Investment

The nature of the interaction can be illustrated by a simple example. Suppose the firms in an industry with closed entry refuse to go into debt at all. Suppose they can sell equities at an imputed cost of 10 per cent. Suppose also that demand for their product grows at 4 per cent per year. Finally, suppose the firms have a policy of paying out 50 per cent of earnings in dividends. Suppose the ratio of capital stock to demand is such that the average return on investment exceeds 10 per cent. The industry's capital is then growing at over 5 per cent. Since demand is growing at only 4 per cent, the ratio of capital stock to demand will rise and the profit rate will fall.

Suppose the average return reaches 8 per cent while the marginal efficiency of investment is above the opportunity-cost level. Then the firm will finance itself and neither lend nor borrow.

Two other cases may be distinguished. If the marginal efficiency of investment reaches the opportunity-cost level while the average rate of return is above 8 per cent, the firm will have excessive funds. It will either adjust its dividend policy or invest in financial assets (more often it may begin to look for previously unexploited opportunities for diversification and integration).

Finally, suppose the rate of growth is high and the yield on equities low. At some ratio of capital stock to demand, retained earnings will be sufficient to provide the capital required for expansion (without further reduction in capital intensity). But if at that point the marginal efficiency of capital is above the yield on equities, it will pay to issue equities.

So far we have considered only cases in which entry is completely closed. The possibility that new firms will enter or that fringe competitors may attempt an expansion may place an upper limit on the return to capital. Suppose the average return permitted by that upper limit is too low to permit the firm to expand as fast as demand, while the marginal return is below the equity yield. Something has to give. The firm may reduce its dividend pay-out, allow its market share to fall, or give up its prejudices against debt.

In practice, of course, policies with respect to debt and dividends are not rigid. In addition, firms in slowly growing industries can diversify or integrate rather than drive the rate of return down to the opportunity-cost level. Finally, all the firms in an industry do not follow the same policy. Some of the consequences of that fact are considered in the discussion of oligopoly in the next chapter.

Subject to those qualifications, the argument of this chapter may help to explain some aspects of corporate financial behavior.

According to the argument given above, the sources of funds used by a firm will vary with (1) the rate of growth of demand for the firm's product, (2) the intensity of competition and the difficulty of entry into the industry, (3) opportunities for diversification and integration, (4) the cyclical and other risks to which the industry is subject, (5) the cost of equity capital (which will, of course, reflect both risks and growth potentials). In addition, there will be considerable variation among firms because managements' attitudes toward dividend policy, debt, and the issue of equities vary.

On the basis of the argument given above we will expect a firm to operate in the inelastic range of its supply-of-funds schedule (i.e., at its debt limit) if two conditions are satisfied. These are: (1) When the marginal efficiency of investment is as high as the equity yield, the average rate of return on capital is high enough to permit the firm to grow as fast as its demand grows without exceeding its debt limit. The yield on industrial equities is sufficiently high to ensure that this condition will be fulfilled for most industrial firms. (2) When the marginal efficiency of investment is as low as the opportunity-cost level (i.e., the cost of using retained earnings for investment rather than dividends or financial assets), the average rate of profit on capital is too low to permit the firm's assets to grow as fast as demand grows without exceeding the debt limit. Or (2*a*) when the marginal efficiency of investment is as low as the prospective rate of return obtainable from diversification, there are enough opportunities for diversification to absorb the firm's excess funds. Excess funds here means the amounts available from retained earnings plus debt after the firm has invested enough to keep the marginal efficiency of investment at the diversification opportunity-cost level.

We might expect that in safe, rapidly growing businesses with relatively cheap equity capital, profit rates (on capital) would be low, and that both equity finance and debt finance would be used. That situation is exemplified by the behavior of the electric utilities.

At the other extreme, consider the case of a group of firms which are highly efficient, are well protected against entry, form a tightly knit oligopoly, and have a moderate rate of growth. Such firms will earn very high average rates of return. They can pay fairly generous dividends and still finance expansion and keep the marginal rate of return down to the opportunity-cost level unless they diversify or integrate. If they are unwilling to borrow for integration or diversification, they will not raise any outside funds. But their failure to raise outside funds is due not to their resistance to taking on debt but to the absence of need for debt. The large automobile companies are more or less in that situation. (Of course they occasionally borrow, but only a very small fraction of capital is raised in

TABLE 2
RELATION OF INTERNAL SOURCES OF FUNDS, NEW INVESTMENT IN PHYSICAL ASSETS BY MANUFACTURING INDUSTRY, 1946–1953

	Internal funds		
	Billions of dollars	Ratio to plant and equipment outlays	Ratio to plant and equipment outlays and increase in value of inventories
All manufacturing corporations	72.6	1.02	.73
Food	6.7	1.07	.76
Textiles	4.3	1.17	.88
Paper	3.4	1.17	.97
Chemicals	7.2	.87	.71
Petroleum	10.0	.64	.58
Rubber	1.4	1.37	.91
Other nondurables	4.4	1.57	.84
Stone, clay, glass	2.4	1.10	.90
Basic and fabricated metals	12.5	1.05	.80
Machinery, excluding electrical	6.3	1.43	.78
Electric machinery	3.2	1.24	.63
Transportation equipment	6.5	1.07	.63
Other durables	4.3	1.26	.70

Source: U.S. Department of Commerce, Office of Business Economics.

that way.) However, it is probable that some adjustments of dividend policy are used, so that marginal returns do not fall to the opportunity-cost level.

Between those extremes, most manufacturing companies seem to be able to finance their expansion from internal sources plus a rather moderate amount of debt. In the manufacturing field, equity capital is used regularly by a very small number of firms which are growing rapidly. But the fact that most manufacturing firms are able to obtain most of their capital from internal sources does not mean that the supply of capital to those firms is highly elastic (when we include imputed costs in the supply price).

Some data on the sources of funds used by manufacturing companies are given in Tables 2 to 5. Table 2 shows that during the

TABLE 3
CORPORATE* MANUFACTURING INVESTMENT AND ITS FINANCING

	Book value outstanding (end of 1945)		Uses and sources of funds, 1946–1953	
	Billions of dollars	Per cent of total for all corporations	Billions of dollars	Per cent of total for all corporations
ASSETS				
Gross fixed assets†	53	38	72	48
Selected current assets	54	57	43	47
Inventories	17	65	27	63
Receivables	14	54	12	28
Liquid assets‡	23	53	5	56
LIABILITIES				
Long-term liabilities and capital	100	42	91	49
Capital stock and capital reserves	65	46	49	60
Depreciation reserves	29	53	28	45
Long-term debt	7	17	14	33
Selected short-term liabilities§	17	49	19	43

* Excluding banks and insurance companies.

† Exclusive of land and before depreciation.

‡ Cash, deposits, and U.S. government securities.

§ Trade payables, bank loans, and Federal tax liabilities.

Source: U.S. Department of Commerce, Office of Business Economics, and Securities and Exchange Commission.

period 1946–1953 manufacturing corporations financed three-quarters of their asset expansion from internal sources. Table 3 indicates that manufacturing corporations accounted for 48 per cent of expansion of gross fixed assets and 47 per cent of expansion in value of inventories by all corporations. They accounted for only 33 per cent of expansion in long-term debt by all corporations and

TABLE 4

SOURCES AND USES OF FUNDS OF MANUFACTURING CORPORATIONS, 1946–1954

(Billions of dollars)

	1946	1947	1948	1949	1950	1951	1952[a]	1953[a]	First half 1953[a]	First half 1954[a]
Total uses	9.1	16.0	13.0	6.1	21.6	22.2	14.5	14.0	7.0	0.0
Plant and equipment	7.0	8.6	9.0	6.9	7.2	10.5	11.5	12.0	5.5	5.0
Increase in other assets:										
Total	2.1	7.4	9.0	−.8	14.4	11.7	3.0	2.0	1.5	−5.0
Inventories	6.2	4.3	2.7	−2.6	5.2	7.8	1.5	2.0	1.5	−2.0
Receivables	.1	2.5	1.0	−1.0	5.6	2.2	2.0	−1.0	1.0	.5
Cash, deposits, and U.S. Government securities[b]	−4.2	.6	.3	2.8	3.6	1.7	−.5	1.0	−1.0	−3.5
Total sources	8.9	15.7	13.1	5.7	21.2	22.2	14.0	13.0	6.5	.0
Retained profits[c]	3.8	6.7	7.3	4.5	7.6	5.7	4.5	5.0	6.0	5.5
Depreciation	2.0	2.4	2.8	3.2	3.5	4.0	4.5	5.5		
Stock issues	1.0	.8	.3	.1	.1	.7	.5	[d]	[d]	[d]
Increase in long-term debt[e]	1.5	1.9	1.8	.4	[d]	3.2	3.5	2.0	1.5	.5
Increase in short-term debt[f]	.6	3.9	.9	−2.5	10.0	8.6	1.0	.5	−1.0	−6.0
Discrepancy	.2	.3	−.1	.4	.4	[d]	.5	.5	.5	.0

[a] Preliminary and rounded to nearest $0.5 billion.
[b] Includes small amount of miscellaneous current assets.
[c] Including depletion.
[d] Less than $50 million, or in the case of 1952–1954 data, less than $250 million.
[e] Includes bonded debt, long-term bank loans, mortgages, and other long-term debt.
[f] Includes bank loans, trade payables, Federal tax liabilities, and miscellaneous current liabilities.

Source: U.S. Department of Commerce, Office of Business Economics, based upon Internal Revenue Service, Securities and Exchange Commission, Federal Trade Commission, and other financial data.

43 per cent of expansion in short-term liabilities. The contrast would be still sharper if manufacturing corporations were contrasted with electric utilities. The relatively small amount of equity finance by manufacturing corporations is indicated by the data in Table 4.

We can account for the observed situation in two ways. (1) We can suppose that firms are willing to borrow (substantially more

TABLE 5
SELECTED FINANCIAL RATIOS FOR MANUFACTURING CORPORATIONS
(Per cent)

Year	Proportion of sales represented by: Liquid assets	Proportion of sales represented by: Profits before taxes	Proportion of sales represented by: Profits after taxes	Profits after taxes to net worth	Long-term debt to invested capital*	Interest payments on debt to funds available for interest payments†
1929	8.4	6.9	6.0	7.9	9.3	12.9
1940	10.5	8.4	5.7	8.5	11.2	6.1
1945	16.2	7.2	2.9	6.2	9.1	4.3
1948	9.9	9.2	5.6	13.0	12.6	3.1
1949	12.1	7.7	4.6	9.4	13.8	4.3
1951	11.0	9.8	4.1	9.7	13.1	3.1
1953	10.3	8.0	3.5	8.3	15.7	4.3
1954, 2d quarter	9.5	7.7	3.9	8.2	15.7	5.1

* Long-term debt plus capital stock and capital reserves.
† Profits before taxes plus interest paid.
Source: U.S. Department of Commerce, Office of Business Economics.

than they now do) to invest at yields only a little above market rates of interest. In that case the rate of investment is governed by the position and slope of the marginal-efficiency-of-investment schedule and the height of the market rate of interest. The limited amount of borrowing done by manufacturing firms is then explained by the relation between retained earnings and the amount of investment which will yield over, say, 4 per cent. (2) We can suppose that the yield required to induce a firm to borrow rises as the ratio of its debt to its earnings rises. In that case the rate of investment in any year is considerably less than the amount which would yield more than the market rate of interest.

On the first hypothesis, a rightward shift in the marginal-efficiency-of-investment schedule will cause an increase in investment equal in amount to the shift at a given rate of return. On the second hypothesis, the increase in investment will be smaller than the shift at a given interest rate. We argued at the beginning of this chapter that the observed investment behavior corresponds more closely to the second hypothesis than to the first.

On the first hypothesis there is no reason why utility companies should behave very differently from others, but on the second hypothesis there is a reason. Utility financing has a very different pattern from manufacturing finance. Moreover, we should expect on the second hypothesis that utilities investment would follow the acceleration-principle pattern more closely than manufacturing investment. That is what happens.

Paul Clark's[12] study of investment in the telephone industry indicates that investment in that industry can be explained by reference to a somewhat generalized version of the acceleration principle.

The results of the correlation studies cited at the beginning of the chapter are at least consistent with the hypothesis that limitations on the supply of finance play an important role in determining the rate of investment.

Finally, statements by business executives indicate a marked reluctance to take on more than a limited amount of debt.[13]

VI. SUMMARY AND CONCLUSIONS

Both theoretical and empirical arguments support the hypothesis that for many firms the effective supply-of-capital schedule (including imputed costs in supply price) is much less than perfectly elastic to the rate of return. That is not due (in the case of large firms) to increases in rates of interest as the amount borrowed rises.

[12] Paul Clark, "The Telephone Industry," in W. W. Leontief (ed.), *Studies in the Structure of the American Economy*, Oxford University Press, New York, 1953.

[13] Charles Silberman, "How Much Can Business Borrow?" *Fortune*, June, 1956.

Rather it is due to increases in the risk premiums imputed by firms as the amount of debt rises.

At the end of the last chapter it was shown that the schedule of marginal efficiency of investment for any one year is downward-sloping. The rate of investment is therefore determined by the intersection of a rising supply schedule and a falling demand schedule. Each of those curves summarizes the influence of a number of factors.

It is clear that one cannot bring all the elements of the theory of investment into a single simple formula. At best, we can summarize the theory in a few propositions, each of which is subject to numerous qualifications. We have argued that:

1. The marginal-efficiency-of-investment schedule is fairly elastic over the range of single-year rates of investment, which is relevant.

2. The primary determinant of the level of the marginal-efficiency schedule (for investment in existing product lines) is the ratio of capital stock to a measure of demand. It will also be influenced by expectational factors, some of which are themselves influenced by the pattern of past movements of income. Prices of capital goods, and other inputs, and rates of technical change will also be influential.

3. The supply schedule of investable funds will vary in level and shape with the debt position of firms. For public utilities it will start at a fairly high level but have only a moderate slope. For many, but not all, manufacturing companies the short-run supply schedule will have an elastic range and then a steeply rising range. The width of the elastic range in terms of rate of investment will depend on the amount of retained earnings and depreciation expense.

4. Depreciation expense is roughly proportional to capital stock. Retained earnings depend on profits and past dividends. Profits, in turn, depend primarily on the size of capital stock and income, varying positively with income and negatively with capital stock.

5. The rate of investment is determined by the intersection of the marginal-efficiency-of-investment schedule with the supply-of-funds schedule.

6. The amount of funds available at a given supply price depends on capital stock, aggregate income, debt, and past dividends. The

marginal efficiency of investment depends on the rate of investment, level of income, and capital stock (plus the other factors mentioned above).

It follows that the rate of investment will depend on level of income, size of capital stock, debt position, and lagged dividends. But the relations between investment and those variables will be shifted by changes in expectations, rate of change of technique, and relative input prices.

In periods of high prosperity many firms will be operating in the inelastic range of their supply-of-funds schedule. As long as they do so, limited variations in the marginal-efficiency-of-investment schedule, whether due to expectations or other factors, will have little influence on the rate of investment.

VII. SOME IMPLICATIONS

A number of writers have based business-cycle theories on the properties of the multiplier-accelerator process. The analysis of multiplier-accelerator processes is based on two postulates: (1) that the consumption function is independent of the rate of investment, (2) that the rate of investment depends on the rate of change of income. However, it need not be the instantaneous rate of change. Some writers, e.g., Hicks, use a distributed lag in their investment functions. That amounts to saying that the rate of investment depends on a weighted average of past changes in income.

The theory developed here implies that both those postulates must be changed. We have shown that the proportion of income going into profits and retained earnings varies with the factors which determine the rate of investment. The consumption function is not, therefore, independent of the rate of investment.

Acceleration-principle theories imply that the rate of net investment will vary in proportion to the rate of change of income. (If a distributed lag is used, that statement applies to the average rate of change of income over a number of years.) The argument of the last two chapters shows that net investment will tend to vary less than in proportion to rate of change of income. The basic reason is very simple. If the ratio of capital to income remains constant, the ratio of profits to income and the ratio of retained earnings to income remain constant. If the rate of growth of income rises and the ratio of capital stock to income remains constant, the ratio of investment to income rises. When the ratio of investment to income rises while the ratio of retained earnings to income remains constant, the ratio of debt to income and of debt to profits must rise. That will not be justified unless the marginal efficiency of investment rises. The marginal efficiency of investment will not rise unless the ratio of capital to income falls. Therefore, as the rate of increase of income rises the ratio of capital stock to income must fall until a new balance is struck between the increased marginal efficiency of investment and the increased imputed cost of debt.

That argument is weakened by the fact that retained earnings tend to rise with the growth of income because of the dividend-adjustment lag. But that consideration only serves to increase the connection between the rate of investment and the propensity to consume.

It is also true that an increase in the rate of growth of income will tend to raise investment by improving expectations. The improvement in expectations will raise the marginal efficiency of investment for given ratio of capital stock to income. Better expectations also make firms willing to increase their indebtedness. But there is no reason to suppose that investment grows in proportion to the rate of change of income as a result of the connection between expectations and rates of change.

The effect of these considerations is to reduce the chance that the economic system can generate very high rates of growth of income. At the same time they reduce the sensitivity of income movements to shifts in the investment or consumption function. The damping effects of "capital rationing" and a high correlation between the propensity to save and the rate of investment reduce the tendency for the economic system to develop strong cumulative processes in either an upward or a downward direction. Those points are developed more fully in Chapters 10 and 11.

CHAPTER 6

Oligopoly and Investment

For reasons of exposition the theory developed in the last chapter was based on the assumption that all firms are pure monopolies and have no fear of new entrants into their territories. In fact, of course, there are relatively few pure monopolies and a great many oligopolies. It is necessary, therefore, to take up the vexing topic of the theory of the firm under oligopolistic conditions. In this chapter we shall begin with the static theory of price for oligopoly. It is argued below that the typical relation of prices to costs under oligopolistic conditions is not very different from the relation that would exist under monopolistic conditions. However, it is also shown that each firm in a small group of close rivals has an especially strong incentive to avoid letting his costs rise much above those of his rivals. There are also some reasons for believing that cost reductions are more likely to lead to product improvement than to price reduction. Both of those considerations tend to speed up obsolescence and therefore increase both depreciation allowances and gross investment.

The possibility that new firms may enter an industry also has some important effects on investment behavior.

I. THE PROBLEM OF OLIGOPOLY PRICING

The problem of oligopoly arises when the number of firms in an industry is so small that each firm must take account of the reactions of its rivals in formulating its price policy. On that definition a very large proportion of the firms in the economy are members of oligopolistic groups. There are a few sectors like agriculture in which no element of oligopoly is present. But in most industries each firm finds that its price policy has a substantial impact on the sales (and thereby on the price policy) of a few other firms which produce similar products and sell in the same geographical areas.

Since oligopoly is so common, it appears in industries which differ from one another in a great many respects that are relevant to price determination. The products of the firms in an oligopolistic group may be standardized or differentiated. The barriers to entry into the industry may be high or low. "Fringe" competitors may play an important role in a variety of ways or be altogether absent. The relations between manufacturers and dealers vary widely and influence price policy in a number of ways. There may be special devices for stabilizing prices and sales, e.g., patent licensing agreements or uniform cost accounting.

In view of these considerations we cannot expect to have a simple theory of price under oligopoly which covers all the industrial structures in which oligopoly exists. In many cases it will be found that price policies are governed by some special consideration which is more or less limited to the industry in question. Nonetheless, in every industry in which oligopoly is present the classic problem of oligopolistic pricing is present and must be solved in some way. If each firm in the group knows that the prices charged by its rival depend on its price, how can the prices be determined? It is argued below that it is possible to make a generalization on this problem which will cover the concrete solutions reached in particular industries as special cases.

II. THEORIES OF OLIGOPOLY PRICING

The oligopoly problem was discussed in a somewhat desultory fashion before 1930. Serious discussion of the problem really begins with Professor Chamberlin's *Theory of Monopolistic Competition.*[1] Chamberlin offered a solution for the problem in the special case in which the two firms are identical in all respects. Later on, the so-called "kinked-demand" solution, developed by Sweezy and by

[1] E. H. Chamberlin, *The Theory of Monopolistic Competition*, 8th ed., Harvard University Press, Cambridge, 1956.

Hall and Hitch, had a certain popularity.[2] That solution, however, requires us to assume that the producers are unable to communicate with one another even when it is to their interest to do so. In addition, it leaves out of account the possibility that one producer can use the threat of price war or other aggressive tactics to force the other to follow a certain price policy.

That is a possibility which always exists in oligopolistic situations. Professor Fellner[3] has recently reviewed the whole literature of oligopoly theory. His own solution meets the objections which have been raised against the earlier proposals. It does not depend on the assumption that the firms are identical. Nor does it depend on any dubious assumptions about what A thinks B will do if A takes a certain action.

To put it very briefly, Fellner begins with the view that the profits of a group of oligopolists will always be maximized if they form a cartel and bargain about the distribution of the cartel profits. It follows that the division of profits can always be arranged in such a way that the profits of any member of the cartel will be greater than that member would obtain under any other method of operation. That fact does not provide a basis for distribution of the profits. The distribution problem can be solved only (*a*) by actual price warfare, (*b*) by agreement on what the outcome would be if it took place, or (*c*) by finding a distribution such that each party agrees that the possibilities of loss as a result of carrying on a price war outweigh the possibilities of gain.

In fact, of course, formal cartels are illegal in the United States and may be difficult to form in any case because managements are unwilling to give up their independence. Inability to allocate output to the most efficient plants or literally pool and distribute the earnings of the group makes it impossible to choose the prices which maximize the group profits and then distribute the profits.

[2] See R. L. Hall and C. J. Hitch, "Price Theory and Business Behaviour," *Oxford Economic Papers*, no. 2, Oxford University Press, London, 1939; Paul M. Sweezy, "Demand under Conditions of Oligopoly," *Journal of Political Economy*, vol. 47, August, 1939.

[3] William J. Fellner, *Competition among the Few*, Alfred A. Knopf, Inc., New York, 1949.

The aggregate profits and their distribution are both determined by the choice of a set of prices or outputs for the different firms. Those limitations do not change the fundamental nature of the problem. If each firm must operate separately, the firms cannot maximize their joint profits, but they can confine themselves to those sets of prices which have the property that for given profits for one firm the others are maximized. They then have to bargain about which of the sets of prices satisfying that condition are to be adopted. The difference between separate and joint maximization is perfectly analogous to the difference between the results of the new and old welfare economics.

If we confine ourselves to the case in which the firms must operate separately, there are two problems: (1) the relation of the shape of the profit "contract line" to the costs of the firms and the demand functions for their products and (2) the factors which determine at what point on the contract line a bargain can be struck and whether a bargain can be struck without a price war.

It is fairly clear that bargaining power depends on relative costs, financial resources, and control of such strategic factors as patents, raw-material resources, transportation and distribution facilities, and related products (not to mention aggressive management and good bargaining technique).

As to the contract line, its slope will clearly depend on the relative levels of the cost curves of the firms, on the scale of their plant, and on the sensitivity of consumers to price differences among the products.

Given sufficient information about the objective situation, risk preferences, and bargaining techniques of the parties, it might be possible to work out solutions to the bargaining problem inherent in the situation just described. We need not attempt to solve that problem, however, because there are additional factors in the situation which make the outcome much less problematical.

III. AN ALTERNATIVE APPROACH

The analysis of the oligopoly problem just outlined has the peculiarity that it takes no account of the past history of the industry.

It is, in fact, almost impossible for the type of situation usually described as oligopoly to arise without a somewhat lengthy historical development. Historically, industrial oligopolies have arisen in a number of ways. In a few instances a monopoly has been broken up into several large firms, which then constitute an oligopoly. In more numerous cases, mergers and the effects of differential efficiency have converted a competitive industry or a large group of monopolistic competitors into an oligopoly. In still other cases, improvements in transport have brought local monopolies into contact in such a way as to create an oligopoly. With new products one oligopoly frequently breeds another, as when two firms established in one field begin to develop a new, related field. For many new products there is a period in which a large number of firms enter the field, there is rapid improvement in the product, a rapid fall in costs, after which there is a series of mergers and failures and the industry begins to settle down. One can hardly apply the usual oligopoly analysis to this sort of development period. Finally, a monopoly is sometimes replaced by an oligopoly either when a new process is developed or when temporary circumstances, such as a war, prevent the monopoly from meeting the demand.

In any of these cases we have to begin our analysis with the assumption that, before conditions which can properly be described as oligopolistic arose, each firm has already experienced a certain volume of sales, has certain customers attached to it (of which more anon), has built up a dealer organization, and has a plant more or less adapted to its sales experience. For reasons which are set forth in the next section, every firm will feel that it cannot afford to accept, without a struggle, a structure of relative price and product dimensions which would cause its market position in terms of share of total sales to deteriorate significantly. At the same time it will be shown that firms will attach a higher value to the maintenance of their existing market position than to improving it at the expense of others.

A. Customer Attachment

It is usual, in discussing demand, to assume that the sales of a product depend only on the price and product dimensions of that product and of its rivals. It follows that, if the price of one product is raised (other prices constant) and then brought back to the old level, sales of the product will first decline and then rise again to the old level. In reality that is seldom true. Generally speaking, a firm which loses sales to its rival by raising prices cannot get all of them back by reducing the price to the old level. For any two products which are at all similar there will always be (for any given set of prices) a certain number of customers who are indifferent or nearly indifferent as to which of the two products they buy. They have to make some decision, and in the first instance of purchase their choice will be determined by random circumstances. But having once purchased one of the rival products, they are more likely to keep on buying it than to switch to another variety. A certain number of buyers will, of course, feel that they did not get what they bargained for. But most of those who are reasonably well satisfied with their decision will see no reason to change.

In the case of industrial customers, there are fairly strong reasons for continuing to buy from the same supplier as long as he gives satisfaction. On the first purchase, credit and delivery arrangements are made, and it takes a certain amount of trouble to change these. Personal relationships may be developed between the salesmen of the suppliers and the staff of the purchaser. Moreover, a continuous relationship with a supplier ensures a buyer of favorable treatment during periods of shortage. In short, a rival seller must offer some positive advantage to induce a buyer to change his source of supply.

Similar considerations apply to some consumers' goods which involve credit, service, or delivery. For most manufactured consumers' goods, however, those considerations are not so important, but others take their place. Of these, the most important is the necessity of making many decisions quickly, which prevents pondering small differences among varieties, together with a repugnance to admitting that one is acting in a random way. Together those considerations tend to make any consumer repeat any choice which does not prove positively unsatisfactory. If the situation remains constant, those repetitions will become habitual, so that buyers almost automatically ask for their "favorite brand" or pick it off the shelf without thinking about the choice at all.

There are, so to speak, two components in decisions to purchase. One is based on the objective qualities of the products and the "real" tastes of the consumers. The other is due to random factors in decisions and to the inertial factors mentioned above. When two products are absolutely identical and are sold at the same price, the distribution of sales depends entirely on random factors and inertia. On the other hand, when products are markedly different, the objective qualities and real tastes of purchasers will be much more important, and random and inertial factors much less important. For, when products are different, there will be only a few customers who are nearly indifferent as between the two products. It is only the nearly indifferent customers whose choice will be much affected by random and inertial factors.

Those two components in decisions to purchase correspond to two senses in which one may say that customer attachment to products is strong or weak. When we speak of customer attachment to cigarette brands, we may say that customer attachment is strong, in the sense that, when the prices of similar brands are fixed, there is a very low probability that a smoker will switch from one brand to another. At the same time, however, it might be said that customer attachment to cigarette brands is weak because a large proportion of smokers will switch brands in response to a small change in price differentials.

In general, it will be true that strong customer attachment in the first sense is associated with weak customer attachment in the second sense and vice versa. That may not always be the case, however. It appears that purchasers of household soaps are very sensitive to price changes but also switch brands frequently when prices of competing brands are equal. There is a random component in their purchases but little inertia.

Let us now consider a case of competition between two very similar products sold at equal prices. Suppose that customers are sensitive to price differences between brands (as they should be if the products are similar) but are strongly attached to a given brand once they start buying it (if its price is no higher than the price of the rival brand). Now suppose that one of the sellers raises his price and the other does not. The high-price seller will lose a large proportion of his customers. What happens when he lowers his price to the level set by his rival? He will recover some of his customers, but it is unlikely that he will get back all those he lost. Many of those who switched brands will not have had any real preference for his brand in the first place and have no reason to switch back. The seller who raised his price can recover his share of the market only by cutting below his rival. Whether he can do that without provoking retaliation is very uncertain.

Those considerations apply to all cases of very similar or identical products. They also apply to cases in which the tastes of consumers are so similar that there is real meaning to the notion of a price difference which compensates for differences in quality.

In the cases to which the argument applies, it is obviously very important to each producer that he should not let his price rise above the price charged by his rival for any appreciable length of time (or, in the case of significant differentiation, should not let his price rise relative to his rival's without a compensatory improvement in quality). Moreover, so long as relative terms offered by the rival producers remain fixed, their shares of the market will tend to remain fixed. Those considerations form the basis of a "solution of the oligopoly problem," but before coming to that we must consider some other aspects of "market shares."

B. Cumulative Effects of Changes in Market Shares

The fact of customer attachment interacts with another type of connection between the sales of today and the demand function of tomorrow. Any loss of customers of significant size directly induced by allowing a rival to give better relative terms than before is likely to have a cumulative effect and bring about further losses. To some extent the ability of a firm to attract customers from a rival depends on the actual size of existing sales. In the first place, products are to some extent self-advertising. Any new buyer or any buyer dissatisfied with the product he has been buying is more likely to think of buying a product with which he is well acquainted because other people use it than he is to think of buying a less familiar one. Second, dealers can carry only a limited number of brands, and they will carry those brands from which they get the highest sales volume. Any brand losing sales will be dropped by some dealers. Some of the customers will go elsewhere, but others will buy another brand carried by the same dealer. (In the case of exclusive dealership the product must generate a certain volume of sales before anyone can take on a dealership in a given area.) Third, for products which have to be serviced, parts will be less available. Fourth, if the loss is a large one, the firm may have to produce fewer models or styles, or pay high costs to keep a full line.

The result is that a 5 per cent drop in sales resulting from an unmatched price reduction may ultimately lead to a larger reduction. The magnitude of this type of effect will, of course, vary enormously from industry to industry.

These considerations have a number of important effects on the nature of the problem. In the first place, one cannot discuss the relation of price and product dimensions to sales without taking into account the existing distribution of sales. When firms produce identical products and charge the same prices, the division of sales is determined by past sales, not by the results of bargaining (unless one firm can bargain with the other to fail to fill orders), as would necessarily be the case if one conceived of the firms as starting from scratch. The result is that the market share which a firm can expect to attain by any amount of expenditure on price or quality warfare depends on its previous position (aside from the effect of past position on resources).

Secondly, the attachment of customers to firms implies that any loss of market share is a permanent loss, in the sense that the expenditure required to get lost customers back is greater than the expenditure required to keep them in the first place. As a result, a firm will be willing to fight to defend its market share even at a time when, from a short-run point of view, the game is not worth the candle. Even if it is operating at such a high output relative to capacity that warfare is not worthwhile in order to keep its share, it cannot take the view that it will give up now and return to fight another day.

The second group of considerations is important because it increases the importance of market shares, as such, as distinguished from the absolute magnitudes with which they are associated.

We now need to introduce two other types of consideration into our problem. The first of these is the significance of the interpretation of motives lying behind actions. When two individuals interact, the response of one of these to an action on the part of another depends not only on the objective character of the action but also on the interpretation which the non-actor puts on the motives of the actor. In the oligopoly situation a price change may be made with one or two motives: (*a*) to change relative prices and, therefore, market shares; (*b*) to change absolute prices with no expectation of ultimately changing relative prices. Suppose firms *A* and *B* have been charging the same price and sharing the market in some way. *B* cuts his price. *A* does not know whether *B* made his move in the expectation that *A* would follow, thus changing absolute prices but not changing market shares, or whether *B*'s move is the first step in a campaign to make *A* take a cut in sales. Unless *A* feels that the price set by *B* is so low that he would rather lose sales than take it, he will cut his price to the same level as *B*. If *B* takes no further action, *A* may reasonably interpret his original move as motivated by a desire to change the absolute price level (without changing relative prices). If *B* makes a further cut, then *A* must regard this as a declaration of war. Price increases can always be interpreted as designed to raise absolute prices (since no one wants to reduce his market share).

The interpretation of motivation and intentions does not stop there, however. Suppose firm *A* makes a sequence of price moves regarded as aggressive by its rival. The rival *B* might in some cases be prepared to allow his relative price to rise rather than fight a price war. But he will almost always take the view that, if *A* is prepared to risk a price war to gain an advantage once, he will be prepared to do so again if he succeeds. *B* will therefore feel that, if he is going to fight, he should do so now. This view will be reinforced when he considers that A took a gamble on whether B would resist his aggressive action. B will feel that, if he fails to resist, A will take this as a sign of weakness and feel safer in trying again. A's greater tendency to resist as his share falls will be offset by decreased ability to resist. These considerations lead to the conclusion that every firm will have a strong tendency to fight to keep the market share it has. More important though (since there are plenty of reasons for being willing to fight to keep something valuable), they make a firm more willing to defend itself against aggressive action than to take aggressive action, and therefore tend to stabilize the *status quo* in terms of market shares.

One further consideration remains. In the short run there will be a difference between marginal costs for a substantially increased output and marginal costs for decreased output if the firms are operating near capacity. It follows that the firm which gains sales by aggressive action gains less than the losses of the firm losing by such action. This also has the effect of making it more worthwhile to keep from losing a given per cent of the market than to obtain an additional share of the same magnitude.

The effect of the above considerations is: (1) It makes the likelihood of a long-drawn-out conflict as a result of an aggressive move rather high. (2) If the rivals are reasonably well matched, the final outcome is likely to be a draw and (in terms of shares) a return to the *status quo ante bellum*. The return to the previous level may take a long time because of customer resistance to price increases. (3) This makes it likely that, at the beginning of an oligopolistic rivalry, firms will be willing to fight to preserve their historical market shares but a great deal less willing to fight to take away their rival's shares. This state of affairs will be known to all the participants. As a result, they can usually agree without the use of aggressive tactics to maintain such relative prices as will preserve existing market shares or customer relationships. If their products are substantially identical, that implies uniform or nearly uniform prices. If there are quality differences, it implies price differences sufficient to balance them for the existing distribution of customers. In the first case, there is no problem of finding the right relative prices. In the second, some experimentation may be required, but there will usually have been more or less appropriate price differences to use as a base.

C. Absolute Prices and Quantities

If the problem of the adjustment of relative price can be solved in the manner just described, the problem of the determination of absolute magnitudes for price becomes a relatively simple one in most (but not all) cases. Suppose that a group of firms have implicitly agreed to maintain a certain set of relative prices which are expected to preserve market shares. Each firm then has a demand curve based on the assumption of given relative prices which is analogous to that of a simple monopolist. Each firm, therefore, has an optimum position on that demand function. Let us first consider the case in which the product dimensions are taken to be fixed. Then each firm has an optimum price for itself and (on the assumption of fixed price ratios) for its rival. These pairs of prices will not, in general, be consistent. Suppose that for firm A the optimum prices are $P_1 = \bar{P}_1$ and $P_2 = \bar{P}_1$, and for firm B, $P_1 = \bar{P}_1 + S$ and $P_2 = \bar{P}_1 + S$; $S > 0$. If A sets his price at $\bar{P}_1$, B has either to set his at $\bar{P}_1$ or accept a higher relative price than the one which will maintain his market share (unless he is prepared to take aggressive action to force A to raise his price). When the firms have resources for aggression of the same order of magnitude and the differences in the desired levels of prices are not too great, firm B will generally accept the price level set by his rival, and the problem is solved on the "low-man-wins" principle.

If, however, the firm desiring the higher absolute-price level is much stronger than its rival or if the difference in desired level of absolute price is larger, the high-price firm may be prepared to engage in aggressive tactics in order to force its rival to raise his price. Finally, when the firm desiring the high price is not strong enough to force its rival to raise price and there is a large discrepancy in the desired price levels, the high-price firm may find that, in spite of the high cost of taking a reduction in market share, it is better to do that than to stay with its rival. This is, of course, a very dangerous situation for a firm, and the firm is not likely to accept defeat until it has made a determined effort to bring its costs down.

In some cases, the costs of the rival firms may be quite similar and the customers not very sensitive to small price differences. In those cases each firm can set its own price on the assumption that its rivals will charge about the same price. The resulting inconsistency between the actual prices and the "conjectured" prices may be so small that no readjustment is necessary.

In other cases, in which the costs of the major producers are roughly similar, price leadership may appear. So long as costs are similar, firms may find it easier to follow the leader than to set prices and then revise them.

There are exceptions to the "low-man-wins" principle, but they will work only when the desired prices of the different firms (on fixed-shares assumptions) are not very different.

Whenever one firm gains a substantial cost advantage over its rivals, that firm can set the price level for the industry unless its rivals are prepared to engage in aggressive action to raise price.

So long as only downward movements in price are involved, this solution works easily enough. Any firm which wishes to reduce the level of prices can do so by lowering its own price. If any one firm wishes to raise the level of prices but other firms do not, the low-price firms will obviously have their way.

But suppose that all firms wish to raise prices. It used to be argued that every firm would be afraid to raise prices for fear that the other firms would not follow. According to the kinked-demand theory, all firms might wish to raise prices (in response to increased demand or costs) and yet not be able to do so. That seems to carry things too far. It is not necessary for managements to sign an agreement in a smoke-filled room in order to communicate. It is always possible for managers to indicate their willingness to raise prices by giving interviews to the press or by making speeches at the chamber of commerce without violating the law. It is also possible to signal one's willingness to raise prices by raising the prices of extras or minor products to see how others respond. Finally, if a firm believes that other firms are willing to raise prices, it can take the initiative and raise its own prices. In the circumstances the risk that the others will not follow is not great. And the loss from announcing a price increase and then retracting it is not great enough to prevent firms from acting (though it may be great enough to prevent very frequent price adjustments). All in all, it seems unreasonable to suppose that there is any very serious problem of communication.

D. Product Variations

So far we have assumed that, although products may differ from one another, the product dimensions offered by each firm remain fixed. We now have to consider the effects of nonprice competition.

In general, we can suppose that each firm is producing a number of lines at different price levels. It is also possible to sell some quality variation in the form of extras. The most common type of quality variation, therefore, is change in quality with price fixed rather than joint variation in price and quality. Small changes in quality with price fixed are very similar in some respects to reductions in price. Their object is to attract customers, and the increase in production costs required to improve quality corresponds to reduction in unit revenue resulting from a price change. At the same time there are several important differences. (1) The increase in unit cost required to attract a given number of customers (assuming rival product dimensions and prices fixed) may vary widely. Many small improvements in design can be made at little or no cost. Where style is a factor, a given expenditure on a new design will produce highly variable and uncertain results. (2) Product improvements usually require capital outlays, whereas price reductions usually do not. (3) Price cuts can be matched by rivals almost immediately, but it takes some time to copy product improvements. How long a delay is involved depends on the effectiveness of patents on the improvement (or the rules of patent-exchange agreements) and on the length of time required to acquire the technical skills and tools to reproduce the improvement.

The lag in the response of rival firms makes product improvement much more attractive as a competitive weapon than price reduction. Suppose that the elasticity of sales with respect to expenditure on product improvement (in view of the possible improvements available) equals the elasticity of sales with respect to equivalent price reductions (the elasticity in both cases being calculated on the assumption that rival price and product dimensions are fixed). If other prices and product dimensions were really fixed, a firm would be indifferent as between reducing price and improving quality. In fact, rival firms are almost sure to match a price reduction and cannot match the quality improvement for some time. They may make independent quality changes at the same time, but that is another question. The effective elasticity of sales with respect to expenditures on quality improvement is almost certain to be greater than the effective elasticity of sales with respect to price reduction. Much of the gain in sales from the improvement will be lost when rival firms catch up, but some of the gain will be retained indefinitely. If one firm makes a product improvement and other firms follow after a time, the demand situation is essentially the same as it was before the change, except that market shares may have changed slightly. But costs are higher than before. The optimum price (on fixed-shares assumptions) will be higher for every firm than before. Prices generally should go up. However, productivity is generally rising at the same time, so that prices may fail to fall instead of actually going up. In other words, the consumers may get a free ride during the process of product improvement, but they do not get it indefinitely.

As a result of these considerations there will be a strong incentive for firms to compete by product improvement. Every firm will simultaneously attempt to gain sales from its rivals by product improvement (often by introducing different versions of the same thing), and much of the potential change in market shares will cancel out. Out of the set of product changes adopted in any one year, the most successful will be adopted by all firms and the others will be dropped.

Continuous product change will generate a considerable amount of gross investment and correspondingly higher depreciation charges. It is therefore likely to make some contribution to net investment.

E. Entry

We have continued to assume the number of firms to be given independently of the price and output behavior of the existing firms. It is certain, however, that there is some tendency for new firms to enter any field in which high profits are being earned. The "new" firms may be either wholly new ventures or new divisions of firms established in other fields. However, the level of profits which can be earned without attracting entrants varies widely from one industry to another. The size of the minimum investment required to produce efficiently varies widely from industry to industry. Variations in the amount of advertising expenditure and in the length of time required to generate an adequate volume of sales are equally important. Similarly, the amount of investment required to build up a sales and service organization may be just as important as the investment in physical plant.

In industries in which the entry barrier is sufficiently high, the optimum price for the existing firms may be well below the price at which new firms will be induced to enter. In those cases the fear of entry plays no role in explaining observed behavior. There may, however, be cases in which the optimum price for existing firms is above the price which will induce entry. In that case the existing firms may set their price at the level which will prevent entry.[4] The effect of differences in entry barriers is to change the determinants of the "optimum monopoly price" in the solution to the oligopoly pricing problem given above.

When the small group of dominant firms has a fringe of small competitors, the optimum price to the members of the dominant group is again changed. In that case the elasticity of demand to the large firms is influenced by the presence of the fringe competitors. If the dominant firms charge very high prices, the problem of keeping the fringe firms from "chiseling" by giving secret concessions to dealers and other means becomes so great that it is more profitable to charge a lower price.

[4] For a discussion of the significance of entry barriers see J. S. Bain, *Barriers to New Competition*, Harvard University Press, Cambridge, 1956.

More generally, many of the special features of the structure of particular industries can be interpreted as factors which influence the optimum monopoly price to the dominant firms.

F. Some Further Qualifications

Economists have a general tendency to overemphasize price policy when they discuss the theory of the firm. That is so because of long tradition and because of the important role given to price in the theory of optimum resource allocation. This bias on the part of the economists leads to an overestimate of the degree of precision which firms will attempt to achieve in finding the optimum price. It was said above that each firm will have an optimum price on the basis of fixed market shares, and that the firm which wishes to set the lowest level of price can make its decision dominate the decisions of its rivals. That is an overprecise way of stating the matter. What is the optimum monopoly price? At least in times when new plant is being built, it must be above the average cost in the new plants when they are operated at capacity. It must be below the levels at which new firms will enter, at which industrial buyers will make their own supplies, at which fringe competitors will begin to make serious inroads by secret concessions, or below the level at which imports will come in in serious volume.

Whether the optimum price falls at one of those limits or in between depends on the elasticity of demand. In a great many cases, firms seem to feel that market demand is very inelastic. In those cases the optimum price ought to be set at the upper or entry limit. But, except possibly in the case in which imports are important, it is very hard to know where the limit is. Similarly, when the entry limit is very high or demand elasticity obviously significant, it is very hard to know what the elasticity is. The optimum price *cannot* be set with any precision. What firms are likely to do in that case is to set price or margin over costs at a level which yields "fair," i.e., generous, profits on investment but which seems to be safely below the point at which new entrants will appear.

That being done, the management turns its attention to the improvement of technique, to advertising and product improvement with a view to maintaining or increasing its market share. Price may be used as a weapon in the struggle over market shares, but not in the form of open list-price reductions. Discounts and other concessions to dealers which are equivalent to price cuts may be used when the management believes it can gain more than it will lose by retaliation, but the significance of that kind of skirmishing is limited. To put it simply, it does not pay to push the price so near the entry limit that one runs a serious risk of losing part of a highly profitable market. On the other hand, it does pay to make a strong effort to increase market share by nonprice competition.

IV. EFFECT OF OLIGOPOLY ON INVESTMENT

The argument of the last section offers a solution to the problem of the indeterminacy of oligopoly, but it does so only on the assumption that costs are given. Costs depend on technical knowledge and factor prices, but they also depend on the willingness of firms to discard equipment rapidly as new equipment appears and to expand facilities so as to avoid intensive and high-cost utilization of capacity. Over the long run, costs also depend on the willingness of firms to spend money on research.

In a simple monopoly situation, the speed of replacement and the average intensity of utilization of equipment are determined in such a way as to equate the prospective rate of return from further investment to the return required to make the firm willing to take additional risks—as a result of increased debt and fixed charges.

In an oligopolistic situation, there is an additional element. Every firm must be concerned about the relations between its costs and the costs of its rivals. Suppose firm *A* lets its rival *B* obtain a cost advantage. Then *B* may reduce price or increase quality to such an extent that *A* will have either to take very low profits or lose market share. It is important, therefore, for every firm to cut costs as fast as its rivals do. But that can be achieved equally well whether all the firms follow a cautious policy and reduce costs slowly, or adopt a daring policy and reduce costs rapidly. We have only succeeded in moving the indeterminacy from the area of price theory to the area of investment theory. In the case in which each firm knows what the other firms are doing and knows that its own actions can be observed by the others, the situation is essentially the same as in the static case. In those circumstances it will not ordinarily pay any one to step up investment or research expenditures in the hope of getting a cost advantage and then an increase in market share. On the other hand, it will seldom pay any one firm to lag very far behind its rivals. Ordinarily, then, the firm which is willing to take the greatest risks will set the pace of investment and research expenditures which in the long run set the level of costs. That conclusion holds, of course, only on the average. In the short run there will sometimes be wide disagreement about the optimum timing of investment expenditures. Since no one can get much of a cost advantage in a year or two, every firm is free to do as it likes.

That is not true, however, when physical ability to meet demand is in question. Most oligopolistic industries tend to follow a long-run price policy and do not raise prices in periods when there is a sudden increase in the utilization. It can sometimes happen, then, that some or all of the firms in an industry do not have enough physical capacity to meet demand. They can build up order books to a certain extent without fear of losing customers, but if the waiting period becomes very long, construction of new capacity becomes a competitive matter. If any one firm is willing to build additional capacity, other firms must do so also. Otherwise the firm with the new capacity will serve some of their customers, so that they cannot get them back. Since one cannot conceal a whole plant, there is no danger that a large-scale capacity war will occur. But, as in the other cases, the firm with the greatest readiness to build will force other firms to build, though they would not do so if they were pure monopolists.

So far we have discussed cases in which each firm has full knowledge of the other firms' action. Construction of new capacity is hard to conceal, but replacement expenditures and research are relatively easy to conceal. Their results should be reflected in published financial statements, but such statements do not tell anything about individual products, while the vagaries of cost accounting conceal the costs which are relevant for price-setting purposes. For that reason a firm can hope to build up a sizable concealed cost advantage or get ahead of its rivals in the development of new products. If it wishes to pursue an aggressive policy, it has some hope of succeeding. In those circumstances, firms which are not aggressive, i.e., not willing to change their investment policy in the hope of gaining increased market share, must follow a more vigorous cost-cutting policy than they would like. They must do so not because they know their rivals are aggressive but because they fear they might be. Every firm therefore has to steer a course between the cyclical risks of investing too much and the competitive risk of investing too little.

For that reason the investment behavior of oligopolists will be substantially different from that of monopolists. From the point of view of demand generation and cost reduction, an oligopoly is superior to a simple monopoly, even though there is not much difference in the average profit margins obtained in the two cases.

However, the tendency for oligopolies to generate competitive cost cutting may be frustrated in a number of ways. Formal or informal market-sharing agreements obviously weaken the drive for cost cutting. Cost reporting and standardization of cost accounting reduce the chances of concealing cost reductions and therefore the risk that a rival is stealing a march. Any firm which controls key patents or resources can protect itself against price cutting by rivals by withholding licenses or resources. Its fear of aggressive action is reduced, and the possible gains from aggressive action are reduced for the others.

It follows that the investment behavior of an industry depends on a number of institutional features of the industry's organization as well as on the number and size of firms.

It will be noted that there remains an element of indeterminacy in oligopolistic behavior. Each firm's behavior depends on what it thinks other firms may be doing. That, in turn, depends in part on past history. Once the firms in an industry embark on aggressive cost reduction, they will find it difficult to stop. But if they follow a live-and-let-live policy from the beginning, no firm need be aggressive or suspect other firms of being so. Whether or not some firm follows an aggressive policy is partly a question of chance, but it is also partly a question of the general social milieu in which managements live. The behavior of oligopolists cannot, in the end, be completely explained in terms of rational profit maximization.

Subject to the qualifications just made, we may expect to find that oligopolists tend to replace equipment more quickly than do pure monopolists. They do so because they have a greater incentive to reduce costs and improve their products than have pure monopolists. The competitive incentive will increase gross investment and depreciation charges and will therefore influence net investment through the lag between depreciation and replacement. One might also expect that oligopolists would be willing to apply somewhat lower risk discounts to the yields on investment projects than would pure monopolists. That should result in greater capital intensity and a greater willingness to borrow than would be the case in a pure monopoly. Those conclusions do not lead to any serious change in the formal structure of the theory of investment developed in the last chapter. They do imply, however, that the theory of investment at the microlevel must take into account the complex of factors which we lump under the heading "market structure."

CHAPTER 7

Investment in Housing

The theory of investment developed in the last three chapters applies primarily to investment by industrial and public-utility firms. In a crude way it applies to agricultural and mining enterprises. It also has some application to trade and service enterprises. Obviously, many modifications would be required to meet the special circumstances of different types of enterprise. But the broad conclusions hold for all types of business enterprises.

They do not hold at all for investment in residential housing or for investment in general-purpose office and store buildings, yet investment in residential housing has accounted for 25 per cent of gross investment in the period 1950–1956. Commercial construction (including some special-purpose building) accounted for another 5 per cent of gross investment in the same period. Obviously, one cannot have a theory of aggregate investment without taking account of the special factors governing the rate of residential construction.

I. STRUCTURE OF THE HOUSING MARKET

The processes which determine the rate of construction of residetial housing are extremely complex. New construction has to be sold in competition with existing housing, and the market for existing housing is split into many differentiated subsectors by location, type of unit, type of tenure, quality (in many dimensions), and age. Moreover, the markets for owner occupation and rental are extremely imperfect. At the same time new construction is carried on by a great many firms of many different types.

Because of the cost of transport, each city or metropolitan area constitutes a separate housing market. Except in wartime conditions it is probably safe to assume that the population and income of a city or metropolitan area can be regarded as given independently of the state of the housing market in the area. (It is possible, of course, that high land or building costs will force firms in a rapidly growing area to raise wages faster than their rivals in other areas, and that this will check the growth of the area.) In what follows, then, we shall deal with a single geographical housing market, taking its population and income as given.

At a minimum, two sectors of the housing market must be distinguished, rental-housing units and single-family houses for owner occupation. Neither of these two sectors is in any sense homogeneous. Rental units range from old-law tenements to Park Avenue penthouses, and single-family dwellings from the tiniest bungalow to a mansion.

At any given moment, the supply of housing consists of a number of units of each of a variety of types of houses. They can be regarded as making up a joint distribution of numbers of houses by tenure, units per building, location, size of unit, age, condition, and other quality dimensions. Given sufficient time for adjustment, the number of units of each type demanded will depend on (1) the joint distribution of family characteristics in terms of income, age of head, family composition, ethnic background, asset holdings; (2) the structure of rents and house prices relative to other prices and financing conditions for the housing units available.

In the short run, however, most of the units will be occupied, and most of the families will be separately housed. The actual number of units of each type demanded will be greatly influenced by the joint distribution of the characteristics of families and the houses they now occupy. People will not move because of transient changes in income, and it may take a larger difference in rent to induce a family to move from one apartment to another than to induce them to choose one apartment rather than another once they have decided to move.

Moreover, while there is always a rental price for rental accommodation, prices for the great bulk of owner-occupied houses are purely hypothetical. Only a few houses are actually on offer at any given time. The prices which are supposed to influence the demand for owner-occupied housing are therefore the priccs as estimated from current sales or current asking prices. Finally, we do not want to say that the difference between number dcmanded and number supplied of any type of house at any moment is exactly equal to the vacancy rate. We must count a man who offers his house for sale at a realistic figure as a demander of the type of house to which he plans to move when he sells, not as a demander of the type he now occupies.

The inertia and irreversibility of demand of owner-occupiers, coupled with the fact that owner-occupiers usually do not move until they have sold their present houses, accounts in large measure for the extremely low vacancy rates recorded for single-family dwellings during the depression of the 1930s. If it were not for these considerations, price elasticity of demand could not so nearly have adjusted demand to fixed supply without a much greater movement of prices than actually took place.

An equilibrium analysis of housing markets would be extremely complex if it took account of the variations in the houses available and the characteristics of families. Moreover, a very detailed analysis would not be very useful for our purposes. Most of thc detail would have to be suppressed in an effort to integrate the theory of investment in housing with the theory of income generation. Fortunately, there are some reasons for believing that an aggregate treatment of the supply and demand for housing will not lead us very far astray. The aggregation problem is discussed in the Note on the Measurement of Housing Stock at the end of the chapter. However, it is necessary to analyze the markets for rental and owner-occupied housing separately.

We shall analyze housing demand and supply on an aggregate basis. But there are other complexities which cannot be avoided. The rate of housing construction is the result of three interacting dynamic processes: (1) the adjustment of prices and rents to the demand and supply of houses; (2) the adjustment of the rate of construction to rents, house prices, and construction costs; (3) the changes in supply resulting from construction of new houses and deterioration of old ones. In addition, the rate of investment in housing influences the rate of growth of aggregate income, which in turn influences the demand for housing. None of the adjustment processes takes place instantaneously. The lags cannot be neglected because they play an important role in the processes governing the generation of income.

Our procedure in this chapter is as follows. The next section contains some brief comments on the determinants of demand for rental and owner-occupied housing. In sections III A and III B we consider the way in which prices and rents move over time. This analysis can be used in an equilibrium framework, but for many purposes the dynamic aspects of the price-determination process are more important than the equilibrium position. In sections IV and V we consider the factors determining the relation between changes in rents and prices and changes in the rate of construction of single-family houses and apartments. In section VI some attention is given to speculative activity in the apartment-house field. Finally, all these interactions are brought together in a system of equations. The results of the analysis are then applied to a discussion of the so-called "building cycle."

II. THE DEMAND FOR HOUSING

As we have already indicated, the demand for single-family dwellings on the part of a population with a fixed (percentage) distribution of demographic characteristics can be regarded as depending on (1) aggregate income, (2) number of families, (3) nonhousing assets, (4) housing assets, (5) price of houses, (6) level of rents, (7) prices of other goods, (8) credit terms.

The value of housing assets depends on the stock of housing capital and its unit price. A rise in housing prices therefore increases the net assets of housing owners and thus increases their ability to make down payments on better houses. That effect is, of course, swamped by the reduction in the housing value of nonhousing assets held by renters and by the increase in carrying charges resulting from price increases. The asset effect does, however, somewhat reduce the elasticity of demand for housing.

The demand for rental housing depends on the same variables, except that assets are probably of little importance, and the signs of the effects of rents and house prices are, of course, reversed.

The amount of nonhousing assets is important because it determincs the ability of potential buyers to make down payments. The assets which are relevant from this point of view are those in the hands of non-house-owners who have enough income to meet carrying charges on a house. Those who already own houses might, of course, buy better ones, but for the most part mortgage amortization ensures that they have enough assets to buy as expensive a house as they can carry. The down-payment factor is most important in determining when a family can purchase its first house. Generally families that buy a house do so by the time the family head reaches the age of forty and often sooner. The assets which influence house purchases, then, are those in the hands of families formed in, say, the past fifteen years. The amount of assets held by such families is not directly related to the total amount of assets. Total assets will change not only with accumulation but also with capital gains and losses due to changes in valuation factors or to changes in profits as a result of any change in income relative to capital stock. Such changes in capital values will have relatively little effect on the young families saving to buy a house. Their assets are small relative to their incomes and are likely to be held in fixed money forms, e.g., savings accounts, rather than in corporate shares or interest in unincorporated business. The assets available to the largest group of potential house buyers, therefore, depend on the amount they have saved over the last ten or fifteen years. That amount will depend primarily on the average level of income per family over the past ten or fifteen years and on the number of families formed during those years.

The impact of the lagged income variable will depend on the size of the down payments required. When the down payments are large, past income will play a relatively large role and current income a relatively smaller one. Moreover, the length of the past period which is relevant will become larger.

This factor greatly strengthens the influence of credit policy on the effective demand for housing. The interest rate as a cost undoubtedly has some influence, but policies which influence down-payment requirements may be still more important. Down-payment requirements can be influenced in three ways: (1) They are influenced directly by regulation of down payments on mortgages guaranteed by the government. (2) The standards imposed by financial institutions tend to vary with the availability of credit generally. This influence may be particularly strong when the interest rates on guaranteed loans are fixed by the government. Funds for low-down-payment, guaranteed mortgages tend to dry up when bond yields get close to the fixed interest rate on the guaranteed mortgages. (3) With a given general credit policy the volume of funds available for mortgage loans can be varied by government purchases of mortgages in the secondary market.

One more variable needs to be added to the demand equation. To take account of inertia and irreversibility, we can add last year's housing consumption to each of the demand equations. This poses some difficulties of measurement. Consumption of housing services is not demand, since the market may not have been in equilibrium. Nor is it supply, since there may be vacancies. However, for simplicity and because vacancies do not vary so much except in depressions, we can use last year's stock of housing of each type as an estimate of consumption.

The rate of migration may play a role in determining the demand for housing. When families move from one area to another, they increase demand for housing in one area and decrease it in another. One might suppose that when construction is going on in both areas there would be no net result. However, young families may double up with their parents if they stay in the same locality. They cannot do so if they migrate. To the extent that migration reduces doubling-up it increases aggregate demand.

Moreover, part of an increase in demand for housing in older areas may be satisfied by the conversion of large old houses into apartments. If migration holds down the growth of the older areas, conversion may be reduced there while single-house construction speeds up elsewhere.

III. CHANGES IN PRICES AND RENTS

A. Changes in Prices of Single Houses

When the owner-occupier of a single house offers his house for sale (basing his decision on asking prices or on the prices at recent sales of similar houses or on the advice of a broker), he generally offers it at a somewhat higher price than the one he thinks he can get. The house is then listed by one or more brokers and may be advertised. If he is so fortunate as to find a purchaser at the initial asking price, he goes away rejoicing. If not, he reduces the price either by reducing the asking price or by becoming prepared to accept a lower price when offered. The price he asks or is prepared to accept will fall progressively the longer the house is on offer and unsold, until either it is sold or the owner decides to remain where he is rather than take a still lower price.

The asking prices of all similar houses in a given community will move practically together, so that, roughly at least, we can speak of the going supply price. The probability that any given house is sold at a given price depends on the size of the gap between supply and demand at that price. If demand and supply are approximately in balance, most houses will be sold within a fairly short time at a moderate concession off the initial asking price, and there will be no net movement in prices. If supply exceeds demand at the going price, houses will remain a long time on offer and owners will have to reduce the price they ask. New sellers coming in will start with progressively lower prices. Conversely, if demand exceeds supply, houses will be "snapped up" at the initial asking prices, and prices will drift upward. The change in price during any period of a few months will therefore depend on the size of the excess demand or excess supply as a proportion of the total stock of housing.

B. Changes in Rents

An essentially similar argument applies to the determination of rents taken by themselves. Owners of multifamily dwellings are in a competitive position which is roughly like that of the large group in monopolistic competition (but with somewhat peculiar entry conditions which are discussed later).[1] Their marginal-cost curve is horizontal at a very low level up to the number of units they have, after which it is vertical. The demand curve for the individual firm (other prices constant) must be supposed to be very highly elastic over most of its range. Indeed, it is prevented from being virtually perfectly elastic only by the costs of moving and the inertia of tenants. However, as "output" nears full capacity, i.e., when average occupancy reaches something over 95 per cent, the individual demand curve must become much less elastic because of frictions. Since tenants are always moving in appreciable numbers for reasons unconnected with the level of rents, e.g., because of changes in income or family composition, there will always be a certain turnover. If an owner charges rents below those of his rivals, he can reduce this turnover to some extent, but he will never reduce it to zero. To obtain an average level of occupancy much above, say, 95 per cent, he will have to charge rents which are so low relative to those of his rivals as to be sure that a new tenant signs a lease almost as soon as an old one gives notice of intention to vacate. To do that, he will ordinarily have to give very substantial rent concessions. Since rent reductions given to incoming tenants will normally have to be extended to all tenants, the marginal revenue from increasing average occupancy beyond a certain point is likely to become negative.

Because there is a certain chance element in the situation, the owner of an apartment does not know exactly where he is on his demand schedule. If he has very high occupancy, he does not know whether this is because his rents are very low by comparison with others or whether he has just been lucky. He does have a certain amount of information about the rents charged by his competitors and about their vacancy rates, and although this information is inaccurate, it serves for judging at least when major changes have taken place. When an owner's occupancy rate rises above, say, 95 per cent and other owners also seem to be enjoying high occupancy

[1] David M. Blank and Louis Winnick, "The Structure of the Housing Market," *Quarterly Journal of Economics*, vol. 67, 1953.

rates, a rent increase which is small enough not to drive out any appreciable number of tenants but which will slow up the rate at which vacancies are filled is in order. Other landlords are in the same position, and the net loss of tenants will therefore depend on the over-all demand curve rather than on the more elastic individual one. After a time, therefore, another increase will take place, and the process will continue until vacancies have been driven up far enough to discourage further rent increases.

Exactly the reverse process operates when the occupancy rate becomes as low as, say, 90 per cent. At that point, owners will be convinced that it is worthwhile to reduce rents in order to reduce vacancies. For each individual landlord, therefore, there is likely to be a range of occupancy rates which does not lead to rent change in either direction. Since, however, the ranges will differ in position for different owners, it is perhaps more reasonable to say that the rate of change of rents varies continuously with the average vacancy rate. When the movements in demand are gradual, this process should hold the variation in vacancies to a narrow range, but when demand changes rapidly, the fact that most of the effect of each rent change is canceled out by the general movement suggests that two or three steps may be required before vacancies are brought down to a normal level.

It has been pointed out that the owners of very small buildings do not suffer vacancy rates which approximate the average over-all rate (when they are charging the same rents as those on similar apartments).[2] If the over-all vacancy rate is 10 per cent, the owners of large buildings have vacancy rates which average out to 10 per cent over fairly short periods, and actual vacancies at a given moment will not vary too far from 10 per cent. The averaging-out period will be much longer for small buildings than for large ones, and at a given moment the owner of a four-family building has, say, either 100 per cent or 75 per cent occupancy. The financial pressure on the building owner to get an immediate increase in revenue, even at the expense of a longer-run loss, is greater for the small-building owner than for the large-building owner. The

[2] *Ibid.*

former is likely to make the first cut in rent sooner than the latter. But at least a considerable part of the gain in tenants will come from the larger buildings, so that the owners of small buildings cannot, as a group, cut rents very much faster than the owners of large buildings without drawing tenants away from the latter and removing the necessity for further cuts.

Thus far we have said nothing about negotiation with tenants. When incomes are falling, tenants may ask for rent cuts and threaten to move if they do not get them. There is no reason for owners to wait until they have vacancies and then cut rents. If vacancy rates are high generally, owners who do not have vacancies will see that, if their tenants move, they will be hard to replace. Rents may fall, therefore, as a result of falling income once some vacancies have occurred. Tenants, of course, never ask for rent increases, but they can threaten to move if rents are increased, and this may have some effect in slowing down the rate of increase. This may be one reason why rents in smaller buildings do not rise faster than those in large buildings. Tenants in small buildings have a better chance to negotiate, and their threats have more effect because of the greater relative loss resulting from single vacancies in small buildings.

In short, then, it seems reasonable to make the rate of change of rents depend on the vacancy rate, i.e., on the ratio of units demanded to units supplied, and possibly on changes in income as well.

In a static situation, with fixed population and income and fixed stocks of houses for rental and owner occupation, rents and house prices would tend to rise or fall until vacancy rates reached a position at which prices and rents would not change. The short-period equilibrium price-rent structure would then rule. When demand changes very slowly, we can perhaps identify the ruling price structure with the short-period equilibrium structure. But when large changes in housing-market conditions occur over short periods, the actual price level and structure may deviate from the equilibrium for a long time.

In the 1920s rents continued to rise until 1925, although supply of houses must have increased faster than demand after 1922 or 1923.

Similarly, in the 1930s rents fell until 1935, although demand probably rose relative to supply from 1932 onward.

In the next two sections we shall consider the rate of construction of single houses and rental housing. Dynamic processes are at work in determining construction as well as in determining prices and rents. After the discussion of construction we can attempt to put the two sets of dynamic processes together.

IV. SINGLE-HOUSE CONSTRUCTION

The great majority of single houses built for owner occupation are built by so-called speculative builders who erect the houses and then sell them at whatever price they can get.[3] Speculative builders range from individuals who build one house on speculation during their lifetime to firms capable of building a thousand or more houses per year. Most of the firms are very small and own a relatively small amount of fixed equipment. Entry into the industry is fairly easy; so is exit. In slack periods many firms leave the industry altogether, while the remainder operate on a reduced scale, doing maintenance work and building for the account of others. A rise in the price of houses and an increase in turnover leads to the creation of many new firms and the expansion of old ones.

In the housing boom which developed after World War II the number of firms engaged in construction approximately doubled from the beginning of 1945 to the beginning of 1948.[4] Thereafter it increased at a much slower rate—about 25 per cent from the beginning of 1948 to the beginning of 1952. The size distribution of firms did not change greatly, except that the share of total employment of the largest firms with over 1,000 employees fell from 20 to 10 per cent between 1945 and 1948. This was due, of course, to the increase in the share of house building and the decline in the share of large-scale government projects in total construction

[3] Sherman Maisel, "Housebuilding in the San Francisco Bay Area," *Journal of Political Economy*, vol. 59, 1951.

[4] Betty C. Churchill, "Size Characteristics of the Business Population," *Survey of Current Business*, May, 1954.

activity. As might be expected, most of the new firms were small, but the size distribution was kept more or less the same by the growth of existing firms.

An analysis of the growth of building activity has to take account of the determinants of entry and of growth of existing firms, and these two aspects of the problem are, of course, closely interrelated. In general, it can be supposed that any builder will undertake building for speculation whenever the price of existing houses (adjusted for depreciation) exceeds his cost of construction by enough to make the effort worthwhile. But no one can be certain of obtaining the going price for a particular house within a given length of time; and if there is a likelihood that it will take a long time to sell a house, there is a possibility that conditions will change, bringing down the whole price level, so that the price received will be lower than the going price at the start of construction. The likelihood that the price received will be below the average level at the start of construction depends on the speed with which houses can be sold and on the rate of change of price.

As we have already shown, the rate of change of price depends on the speed with which houses can be sold, which in turn depends on the balance between supply and demand at the going price. When prices are near the level of costs of building for a particular builder, he will not be willing to risk all his capital unless houses appear to be moving off the market fairly quickly, i.e., unless there is a high ratio of demand to supply at the going price. He can be induced to risk a larger share of his capital either by a rise in prices relative to costs or by a rise in the ratio of demand to supply. The latter factor will come into the picture in another way. The number of houses which a builder can erect per year depends on how much capital he has at risk and on the number of times per year it is turned over, which, of course, depends on the ratio of demand to supply at the going price. Finally, builders provide only a small part of the capital required to build a house, the rest being provided by financial institutions on construction loans. When there is a high ratio of demand to supply, financial institutions are willing to advance a greater proportion of the costs of building, thus increasing the amount of construction which a builder willing to risk a given capital can undertake.

When there is a fixed number of firms in the industry with a given amount of capital, the rate of construction will depend on (*a*) the ratio of existing prices to construction cost and (*b*) the ratio of demand to supply at the going price. For a given number of firms with a given capital, there is an upper limit to the rate of building. Even if all houses can be sold before completion, turnover cannot be greater than the inverse of the fraction a year required to build a house, and there is an upper limit to the proportion of construction cost which financial institutions will advance. (This limit may not be reached, however, because of the limitations on the executive capacity of builders.)[5]

When there is a sudden rise in demand for building, the industry may be brought up against this limit and be unable to expand the rate of construction beyond a certain point no matter how high prices go or how great the excess demand. The capital-supply limitation interacts with limitations on the rate of building imposed by the availability of labor and materials. In certain circumstances the capital-supply limitation may be the ruling determinant of the rate of building. If, however, builders are willing and able to build at a rate requiring more materials than can be produced, materials supplies may limit construction. In a sense, however, the supply of materials will influence the rate of building through its effect on the rate of turnover. A shortage of materials shows itself in slow and erratic deliveries, which lengthens the construction period and reduces the capacity of builders' capital (beside raising their costs). Since prices of materials and labor are generally rising at the same time, and rising faster than the prices of houses, the delays also increase builders' risks, which at least prevents financial institutions from increasing the leverage of builders' capital.

Leaving aside the availability of materials and labor for the moment, we have a supply curve for building in which the supply at a given moment depends on the assets of builders, the level of prices

[5] Fred Balderston, "Scale of Output and Internal Organization of the Firm," *Quarterly Journal of Economics*, vol. 69, 1955.

relative to costs, and the ratio of demand to supply at the going price. When building is profitable, the assets of builders will increase as they plow back the profits from their operations. The rate of profit earned on a given capital depends on margins and turnover; and the proportion of earnings which builders are willing to plow back presumably depends in large measure on the rate of profit being earned (marginal and average profits being more nearly the same in this than in most cases). Thus the rate of expansion of capital and building activity from internal sources will depend on the level of prices (relative to costs) and the balance between demand and supply at those prices. A very considerable part of the expansion of building activity during a building boom does result from the expansion of the rate of activity of firms already in the industry.

But a very great part of the expansion of building results from an influx of capital and management from outside the industry. Existing firms achieve part of their expansion by taking in additional partners or by incorporating and selling stock. Of still greater importance is the very large number of new small firms that come into the building industry under favorable conditions. A very high proportion of the owners of small firms are building-trades workers who continue to work at their trade while constructing houses for their own account.

The balance between the numbers entering and leaving the industry clearly depends on the level of returns to be earned in the industry, and, since some assets are required to get into the business, it must depend also on the value of total assets and on the way in which they are distributed. As in other cases, the maximum number of firms in the industry under given conditions will not be reached all at once. If conditions improve suddenly, there will be a high initial rate of entry because there is a backlog of persons willing and able to enter under these conditions. Since, however, the conditions are not uniform and are not uniformly perceived, it will take some time before all those who were potentially ready to enter at the time when conditions improve will actually do so. The entry rate of this initial backlog will start, therefore, at a high level and gradually taper off to zero. But aside from backlog considerations, in each year a certain number of men will acquire the necessary experience and capital to enter (with fixed population, total assets, and returns in the industry), so that once any backlog is worked off there will be a steady flow of new entrants.

The rate of entry may therefore be considered to vary positively with the level of income, asset valuation factors, population, and returns in the industry, and negatively with the number of firms already in the industry. The exit rate will vary with the same factors but, of course, in the opposite direction.

At a given moment, then, the rate of construction of single houses depends on the number of builders in operation, the size distribution of their assets, the prices of houses, costs of building, and the balance between supply of and demand for houses at the existing prices. Prices and the balance of supply and demand are almost certain to change over time, but even if they do not, the number of builders and the size of their assets will increase for a number of years during a housing boom of the kind which occurred after each of the two world wars.

Thus the observed changes in the rate of building may be due to changes in prices or costs or the ease with which houses can be sold, but they may also be due to entry into the industry and accumulation of assets in the industry.

V. CONSTRUCTION FOR RENTAL

Construction of housing for rental is, if anything, an even more complicated matter than the construction of houses for owner occupancy. The structures in question range from small single houses costing a few thousand dollars to large blocks of apartments costing millions of dollars. Roughly speaking, housing units built for rental fall into four classes:

1. Before World War I a considerable number of low-cost single houses were built for rental. Some of these houses were built on speculation and sold to individuals who in turn rented them out. Others were built for the account of those who intended to operate

them as rental properties. A certain number of single houses were also built by companies, particularly mining companies. Since World War I there has not been much building of this type. On the other hand, a good many single houses came into the rental market during the depression of the 1930s because the owners hoped that prices would rise at a later date.

2. A large number of two-family houses were built in the years just before and just after World War I. Many of these were built on speculation. In a very large number of cases the owner lives in one half of the house and rents the other half out.

3. Small multifamily buildings built for and operated by individuals.

4. Large buildings owned by corporations formed for the purpose of owning and operating the building or buildings in question. In recent years, insurance companies have begun to operate large buildings on their own account instead of simply providing mortgage funds.

The rate at which buildings of these types are constructed depends on the level of rents, the vacancy rate, the cost of construction, and the availability of finance. To the extent that speculative building is involved, the rate of building is also influenced by the factors which determine the rate of speculative building of single houses for owner occupancy.

The principal difference between rental housing and housing for owner occupation lies in the sources of funds for the two types of capital formation. In both cases, of course, mortgages given to financial institutions play a vital role in financing construction. The equity capital for owner-occupied houses has to be provided by the occupier. In the case of rental housing, the equity capital can be provided by anyone who can raise the funds and is willing to take the risks of ownership. Since the ownership of assets is very heavily concentrated, most of the assets in the system are not available as a basis for the equity in owner-occupied houses.

The equity interest in housing built for rental comes from the private resources of interested individuals, which, for the most part, means that securities are sold or bank loans are raised against other assets. The smaller buildings are often erected by persons who intend to operate them, but many of the larger ones are promoted by professional real-estate dealers whose interest is either in the fees associated with the operation or with the exploitation of an option on a site. The persons supplying equity capital will be converting other assets first into cash and then into buildings. Since the rate of investment per year in apartment buildings is only a small fraction of total assets, the conversion process will not have a very important effect on the prices of other assets. But the yield required to get people to take on more building equities will rise as the value of building equities rises relative to the value of total assets.

Let us suppose that there is some given amount of nonbuilding assets in existence and a given level of rents. The number of units of building equities that people will be prepared to hold will then be given by a negatively inclined schedule of number against price. Persons who have large assets and a strong preference for building equities will be willing to give high prices for small quantities, and so on in the usual way. If the supply of buildings is fixed, the price of building equities will have to equal the demand price for the number of buildings in existence. But if new buildings can be produced at constant cost, there is no reason for anyone to buy a building at a price above cost. No one need pay more than cost for an existing building, and if demand exceeds supply at cost price, no one need sell for less.

If the market for rental property were a perfect one, with every potential buyer aware of the level of rents, building costs, and operating costs, one might expect the rate of building to be one which would eliminate excess demand as quickly as possible, i.e., to be determined either by the gestation period, supply of specialized construction facilities, or supply of mortgage credit. In fact, however, the excess demand calculated on a full-knowledge basis is only potential. A certain number of building-management companies will have full knowledge of the situation and may build for their own account, but most large apartment buildings are "promoted" by professional real-estate dealers whose business it is to communicate the knowledge of the bright prospects in new building to potential investors. The amount of new building they can generate clearly depends on the number of potential investors available. If there are only a few, it will take them a long time to round up the necessary capital for each project, while if there are many, capital can be quickly obtained and one project will quickly succeed another.

Smaller projects will not be so frequently promoted, but a similar factor is at work. The persons who might undertake these ventures are persons with appreciable amounts of uncommitted liquid assets (in the broad sense), e.g., retiring business men who are looking for a new investment or can be contacted by an interested party, such as a real-estate broker. There is obviously a random element in the process of bringing the people who have the assets and the right frame of mind into contact with a concrete opportunity. The actual rate of building undertaken over any short period will therefore normally be only a fraction of the "excess demand" for building calculated on a full-information basis.

The rate of construction of rental properties tends to vary with the level of rents relative to building costs and the vacancy rate in such a way as to keep the prospective value of new buildings equal to their construction cost. When, however, a substantial gap between prospective value and construction cost opens up, the rate at which new building closes the gap is governed by the rate at which the resulting profit opportunities are discovered and exploited, which depends on the efficiency of the real-estate business. Capital supply and resource limitations become important only when the promotion of new projects becomes so successful that a very high rate of building is generated.

In general, the level of rents and the vacancy rate play the same role in determining the rate of construction of rental properties as that played by the level of prices and the rate of sale in the case of owner-occupied housing. The role played by the lag between the appearance of profit opportunities and their exploitation is similar to the down-payment lag and the lag in the entry of speculative builders in the owner-occupation field.

The general logic of the dynamic process is the same in both cases, though the nature of the elements underlying these processes is somewhat different. However, the field of multifamily dwellings opens up possibilities for speculative activity which do not exist in the owner-occupation field.

VI. SPECULATION

The discussion above runs in equilibrium terms on the assumption of fixed reaction parameters. But apartment buildings have frequently become the focus of speculative activity which has had a considerable effect on the amount and timing of construction. Titles to buildings circulate like common stocks, and since apartment buildings are rented out much more readily than single houses, it is on the former rather than on the latter that speculative activity focuses. When rents and occupancy ratios begin to rise after a decline or a period of stability, real-estate brokers will attempt to arrange transactions between relatively pessimistic owners and relatively optimistic buyers. One of the arguments used in bringing about sales is obviously that values are rising so that the buyer will be able to resell in a year or two at a profit.

If income is rising and continues to rise and if construction is not under way, this view will prove correct even if the asset preferences of potential buyers and the relation of current values to expected future values remain the same. But once profits have been taken on this basis, the word will spread that quick "killings" are to be had in real estate. Brokers will find it possible to interest buyers who would not previously have considered a venture in real estate. This will cause values to rise faster than revenues and will bring forward the time when new construction can be undertaken. When values are rising, the expected yield from a speculation for resale is much higher than the investment yield, and equity money will become readily available. Ultimately, then, the rate of building is almost certain to overtake the rate of increase of demand for building space, and (investment) yields will begin to fall. If asset preferences and expectations remained fixed, this would result in a gradual decline in the rate of building, but the same forces which raised valuation factors will now operate in reverse. As soon as it is recognized that yields are falling, buyers will hesitate and sellers (some of whom will be overcommitted if they do not liquidate their holdings quickly) will be eager to sell. The resulting fall in values will be much greater than the fall in yield. This will set up a cumulative process, and values will be carried down until necessitous sales are checked by (1) long-term investors picking up bargains and (2) owners whose assets permit holding on for the long pull.

The decline will be made the more precipitous by the long gestation period for buildings of the type in question and by the tendency of promoters to conceal the difficulty of renting new buildings by granting concessions, so that the pressure of excessive building is temporarily passed off to older buildings which are not so interesting to speculators. Speculation tends to focus on new buildings, partly because promoters can make more money out of new buildings, e.g., through the sale of land on which they hold options, and partly because it is easy to believe that a higher price can be obtained for a completed building full of prosperous tenants than for a hypothetical one so that a profit can be made by getting in on the ground floor.

The crash will therefore be set off not by a gradual decline in earnings of buildings but by the appearance of a completed building of some consequence which cannot be rented even with the usual concessions. Of course, the whole process can be checked at any stage by any event which affects the financial community generally in an adverse way.[6]

This concealment will not only intensify the crash but will also contribute to overbuilding, so that values based on nonspeculative considerations will fall sharply. It will then be some time till building becomes profitable again even on an investment basis.

The amplitude of the cycle which can be produced in this way depends in considerable measure on (1) the amplitude of fluctuations in income and (2) the extent of the operations of other factors which temporarily bring building to a halt, e.g., wars. If income

[6] Homer Hoyt, *One Hundred Years of Land Values in Chicago*, University of Chicago Press, Chicago, 1934.

grows more or less continuously and if building is not artificially checked, the fluctuation in investment yields will be limited. Even in a strong speculative boom the amount of overbuilding, i.e., the excess of space erected over the amount at which a new building can earn a satisfactory investment yield, is hardly likely to exceed two or three years' growth in demand. Investment yields will never fall very far below the level at which it pays to build for investment purposes, and therefore they will never rise very fast. A speculative boom will never gain very great momentum, therefore, before it is checked. Since there are always fluctuations in the objective situation, the cycle will never die out completely, but it would not be important if there were no large swings in income or temporary halts in building for other reasons. It is obvious that a large fluctuation in income will produce a large fluctuation in yields and therefore gives speculation a chance to gain momentum on the upswing. If a war halts building, the postwar rise in values is likely to be very great and may therefore generate a large-scale burst of speculation activity.

When we come to use this discussion of building activity in a complete system, therefore, we shall have to modify our equation for building activity to make a separation between movements along a given set of investment functions and shifts in a parameter of those functions brought about by speculation.

VII. CONSTRUCTION COSTS

In our general discussion it was shown that the rate of building depends on construction costs as well as on prices and rents. Variations in construction costs ought to be brought into an explanation of the rate of construction. It is very difficult to do this in a systematic way.

There is no reason to suppose that the costs of producing construction materials are not essentially constant (with respect to rate of output in the long run). In the long run, then, the materials component of construction cost ought not to vary with the rate of construction. The prices of construction materials (relative to other prices) will vary in the long run with the relative increases in efficiency in the industry producing the materials. It may also depend on the relative strength of the trade unions involved.

In the short run, a rapid increase in the rate of construction is likely to force up the prices of some materials (especially those which are produced by competitive industries). That will slow down the rate of construction by reducing builders' profits and increasing their risks. The check will be only a temporary one. After a time the supply of materials will expand and relative prices will tend to fall.

The situation is somewhat different in the case of labor costs. There is no reason to suppose that the supply curve of construction labor is rising steeply in the long run. One might therefore expect that the labor costs of construction would vary with the strength of the trade unions involved and with the rate of technical progress in the industry. Since the unions are relatively strong and since technical progress is relatively slow, that leads to a secular rise in the relative cost of building. That, of course, tends to offset the effect of population increase and rising income on the growth of demand for housing.

An additional factor may be involved. When the rate of construction rises rapidly, there may be a shortage of construction labor. Trade unions find it easy to negotiate large wage increases. In the course of time additional labor will be available, but it will not enter the industry and force down wages. The trade unions may be able to get an increase in relative wages during the boom and keep it for a long time, if not indefinitely.

VIII. SUMMARY AND CONCLUSIONS

It was shown in the last two sections that, speculation aside, the rate of residential construction depends on (1) rents and house prices relative to other prices, (2) vacancy rates for apartments and the ratio of demand to supply for single houses, (3) the number of speculative builders and the distribution of their assets in the case of single houses, (4) the total amount of assets in the community and the risk preferences of their owners in the case of apartment buildings, (5) the level of building and operating costs.

Even when we suppose that income, population, and housing assets change at steady rates independently of the rate of building, there are three layers of dynamic processes in the residential-construction industry: (1) adjustment of rate of building to given supply-and-demand conditions and given prices and rents, (2) adjustment of prices and rents to given supply-and-demand conditions, (3) change in supply due to the fact that construction is going on.

If income and population grow steadily and costs are fixed, there is an equilibrium position for the industry. In equilibrium the vacancy rate for apartments and the ratio of supply to demand for single houses must be such that prices and rents do not tend to change. That is, prices and rents must be such that at the existing *levels* of income and population and housing stock, supply and demand are nearly in balance. At the same time the levels of prices and rents and the number and size of building firms must be such as to generate a rate of building (in view of the rate of sale implicit in the no-price-change condition) equal to the rate of growth of demand generated by income and population growth. Finally, the level of prices must be such as to maintain a building industry of fixed size if growth of demand is linear, or one growing at a steady rate if the rate of growth of demand is exponential.

If we consider single-house building alone, these equilibrium conditions lead to the following system of eight equations.

(1) $D_t = f(p_t, t)$ — D_t is demand at time t, p_t is price level of houses at time t. The time variable in the demand function represents the effect of growth of income and population.

(2) $p'_t = F\left(\dfrac{D_t}{S_t}\right) = 0$ — p'_t is rate of change of price level, and S_t is supply of houses at time t.

(3) $S'_t = \phi(N_t, p_t, s_t) = D'_t$ — S'_t is rate of change of supply. N_t is capacity of building industry. s_t is speed of sale of houses.

(4) $s_t = \theta\left(\frac{D'_t}{S_t}\right)$ — D'_t is rate of increase of demand.

(5) $N'_t = R(s_t, p_t) = D''_t = 0$ — N'_t is rate of change of building capacity, D''_t is rate of increase of demand.

(6) $p_t = \int_0^t p'_t\, dt + p_0$

(7) $N_t = \int_0^t N'_t\, dt + N_0$

(8) $S_t = \int_0^t S'_t\, dt + S_0$

If rental building is included, a set of similar equations has to be added, and the demand for single houses has to be made a function of rent as well as housing price levels, while the demand for rental quarters is a function of housing price level as well as rents. Since, in fact, demand and population do not change steadily, the equilibrium conditions are of very little use in explaining the observed behavior of the housing industry. If we drop the equilibrium conditions in equations (2), (3), and (5), we have a dynamic system whose movements are ultimately governed by the movement of income and population. But when income and population move erratically, or when the parameters of the system are changed (as a result of changes in tastes or institutional arrangements), the movements of the rate of building may follow a course which is very different from the path of movement of income.

IX. HOUSING CYCLES

The structure of the housing industry is such that there is a certain tendency for cycles in the rate of house building to develop. Whenever excess demand appears, prices will eventually rise high enough to induce a rate of building sufficiently high to eliminate the excess demand. That requires a rate of building in excess of the rate of growth of demand. At the moment when excess demand is eliminated and prices stop rising, the rate of building is higher than the rate of growth of demand, and building will not slow down until enough excess supply has developed to lower prices or reduce the rate of sale of houses. That implies that the rate of construction must again fall to a level below the rate of growth of demand until excess demand has again developed, so that the cycle repeats.

If income and population rise steadily and if no speculation develops, the cycle is likely to damp quickly. In the single-house field slow sales and even slowly falling prices can bring about a quick contraction in house building so that only a very limited excess supply of single houses can be developed. The rate of building will therefore adapt itself to the rate of growth of demand very quickly so that the second cycle generated by an initial disequilibrium will be fairly weak. A somewhat greater excess supply can develop in the rental field because of the longer gestation period of large buildings and because a large part of the increase in vacancies will be borne by old buildings. But in all periods except the 1920s, apartment buildings have formed a small fraction of total construction, so that excessive construction in that field can have had only a moderate influence on residential building as a whole.

For periods other than the 1920s we might expect, then, (1) that any initial disequilibrium in the housing industry (whether due to income fluctuation or some other cause) would lead to one major fluctuation in housing construction with relatively minor cycles thereafter and (2) that any fluctuation in income would cause a cycle in house building with somewhat different timing. Those expectations seem to be borne out by the historical record.

A great deal has been written on the subject of building cycles and housing cycles.[7] There is said to be a long cycle in building in general and in residential building in particular which has a longer duration than the general business cycles as defined by the National Bureau of Economic Research. Indeed, the housing cycle has a

[7] J. B. Derksen, "Long Cycles in Residential Building: An Explanation," *Econometrica*, vol. 8, 1940.

W. H. Newman, "The Building Industry and Business Cycles," *The Journal of Business of the University of Chicago*, July, 1935.

Clarence Long, *Building Cycles and the Theory of Investment*, Princeton University Press, Princeton, N. J., 1940

J. R. Riggleman, "Building Cycles in the United States," *Journal of the American Statistical Association*, vol. 28, 1933.

longer period than so-called major cycles. Between the Civil War and 1932 there were four full cycles in housing and six so-called major business cycles (counting 1865 arbitrarily as a trough in both cases). Moreover, the timing of the four housing cycles did not coincide with that of the general cycle. The housing cycle was on the upswing during the whole of two short major depressions, 1882–1885 and 1920–1921. The two cycles reached peaks at roughly the same time in the early 1890s, but the housing cycle did not reach its trough until 1900, several years after the trough of the general cycle. The housing cycle shows a peak in 1905 and a trough in 1918, neither of which corresponds even nearly to movements in the general cycle. The peaks and troughs of the two cycles are nearly together in the 1870s and in the cycle of the 1920s and 1930s.[8]

We have already seen (1) that an initial disequilibrium in the housing industry will produce a cycle which tends to damp out fairly rapidly unless there is some other source of disturbance in the economy and (2) that the housing industry has reserves of demand so that the effects of general depression on the housing industry tend to be spread out over time. This tendency also helps to limit the severity of general depressions.

Suppose the housing stock is below equilibrium for the existing income and population, and suppose the housing industry is building at a rate below the rate of growth of demand generated by income and population growth during a period of general prosperity. In those circumstances, the housing industry can ride through a relatively mild or short-lived general depression without suffering any serious ill effects. It will be affected, but the effect of the general depression will be to reduce the rate of growth of building to a level below what it would have been had the depression not occurred. We cannot measure that kind of effect. The general depression will, of course, be made less severe by the continued growth of building. The situations in 1882–1885 and 1920–1921 correspond to the one just described. In the 1880s a strong housing boom was under way, stimulated by the rapid growth of income and the recovery of im-

[8] David Winnick, *The Volume of Residential Construction 1889–1950*, National Bureau of Economic Research, Technical Papers, no. 9, New York, 1954.

migration after 1879. Both demand and supply lagged behind the growth of income, so that they continued to grow through the mild depression—which presumably resulted from the slowing-down of railroad construction. In 1920–1921 the building industry was faced with a housing shortage and had no reason to contract its operations.

In 1907–1908 the housing industry was somewhat more responsive to the general decline in activity. Demand for housing was still growing in a lagged response to the rapid growth of income and population over the preceding ten years. But the rate of building was probably already higher than the growth of demand. Building activity had declined from 1905 to 1906, and it fell again from 1906 to 1907, but it continued at high levels in spite of the depression and financial crisis. In this case, the building industry had a reserve of demand, but not nearly so large a one as in the other two cases cited.

The situation in the 1890s presents some additional features. The depression of the 1890s was much more severe than that of the 1880s and lasted longer than the depressions of 1907–1908 and 1920–1921. Immigration and, presumably, population growth were depressed, and immigration remained at relatively low levels for several years after the end of the depression. We can only guess the condition of the housing industry. Building activity in 1891 and 1892 was somewhat below the highest levels achieved during the late 1880s, so that there was no shortage of building capacity. Building had proceeded for the preceding few years at reasonably steady rates with no marked trend. There was probably no great oversupply of housing relative to demand.

During the depression of the 1890s, building of houses fluctuated on a mild downward trend but continued to decline until 1900, three years after income had recovered to predepression levels. Building did not reach the level of 1892 until 1905. The fact that building fell off by only 10 per cent during the depression but was still 15 per cent below the predepression level in 1902 requires explanation. (In 1900, building was 25 per cent below the level of 1892, but some special factors were in operation in that year.)

The explanation lies, in part at least, in the reserves of demand backlogged by credit restrictions during the years before 1892, when population and incomes were rising rapidly. Real income fell only a little during the depression of the 1890s. Many individuals who had the income to carry better housing were saving during the 1880s and early 1890s, and continued to do so during the depression. Those who achieved their goal were able to buy houses during the depression. Indeed, the decline in prices increased the real value of their assets and enabled an especially large number of persons to acquire the real assets necessary for home purchase. As time passed, the influence of the depression years on the asset position of house buyers became greater, and demand slowly declined. The decline was accelerated by the reversal of the price movement in 1897. Income began to rise in 1897, but it did not have as great an effect on housing demand as usual because of the slow response of immigration. Finally, the housing industry is slow in adapting itself to excess demand, so that the stock of houses was somewhat excessive by the time the increase in income was able to offset the declining backlog of housing demand.

The housing depression in the 1890s can be regarded as a consequence of the general depression. However, the factors just mentioned spread out the housing depression over a longer period. As a result, the depression was somewhat milder than it would have been (with a more rapid response on the part of the housing industry).

Once the rate of house building had declined, it took a little while to build it up again, and some backlog of demand was again built up. Building increased rapidly in 1904 and 1905, overshot the mark somewhat, and then settled down to a fairly steady rate, with relatively mild fluctuations until 1916. Thus the induced cycle caused by the low level of building activity at the turn of the century was a very mild one.

The trough of building during 1917 and 1918 was clearly due to the entry of the United States into the World War. It has no significance except for the assistance it gave in creating the housing shortage after World War I.

It seems reasonable to take the view that in the period from 1879 to 1916 house building followed the movements of aggregate income and population, but followed them with a response of the distributed-lag type. The tendency of the industry to produce a cycle of its own, independent of the general cycle, seems to have been very weak. The housing industry could ride through a mild general depression or spread out the effects of a severe one. It did not produce any housing depressions (of any magnitude) through its own dynamic mechanism. On the whole, the structure of the housing industry contributed to the stability of the system against depressions stemming from other causes.

But that cannot be said of the period after 1916. The housing shortage which appeared at the end of the war caused a building boom. The boom was enhanced by a reduction in down-payment requirements for purchases of houses. Such a reduction tends to cause a "one-shot" increase in house building because it brings forward the date at which a family can afford to purchase a house. As we should expect, the building boom got under way slowly, and construction did not reach its peak until 1926. This was presumably due to the operation of the various lags discussed earlier. Both house prices and rents rose until 1925 or 1926 and then fell slowly. There was a speculative boom in apartment building in large cities which undoubtedly led to an appreciable amount of overbuilding.

By 1928 residential building was only 13 per cent below the rate achieved in 1926, but in 1929 the residential construction rate was only 60 per cent of the 1926 rate, and in 1930 it was only one-third of the 1926 rate. The rapid increase in construction in the years 1919 to 1926 was the result of the low level of building during the war years (one should not, of course, exaggerate the influence of the wartime stoppage, since it lasted only a couple of years), the increase in income during the 1920s, the easing of mortgage credit, and the speculative boom in apartment building.

Income might have kept on increasing, but the other three factors were temporary. Building could not have continued at the rate achieved in 1925 and 1926. There was bound to be a fluctuation in the rate of building, since temporary factors drove it above the rate which could be sustained by the rate of growth of income and population obtaining in the 1920s. At the very least the rate of building had to fall from the levels attained in 1926 to a level consistent with the rate of growth of income and population. It had to fall still further because the overbuilding of apartments had not only to cease but to be balanced by underbuilding.

Even if income had continued to grow after 1929, residential building would have fallen to a low level for a time and then, no doubt, overshot the mark on the other side. With the aid of a strong speculative boom in apartment building it is possible to get a housing cycle which does not damp out completely after the first upswing and downswing even when income rises steadily. Even then, it seems likely that without cycles in income the cycle generated by speculative building would damp out fairly quickly.

But, of course, the movements of income are not independent of movements in the rate of construction. The rate of building reached a very high level partly because it helped to increase the rate of growth of aggregate income. By the same token, the decline in the rate of growth of building tended to reduce the rate of growth of income, and that required a still further reduction in the rate of building.

We must think of two types of housing cycle: (1) Cycles in which the normal relations between housing investment and income movements operate and in which housing investment tends to follow the path of movement of aggregate income, though with different timing. In this case, there is not only very little tendency for residential construction to move cyclically by itself, but the characteristics of the housing industry are such that residential construction tends to stabilize the system against other sources of fluctuation. (2) Cycles in which changes in the structure of the housing industry itself, such as wartime construction backlogs and speculation, tend to produce a fluctuation in the rate of house building even when aggregate income is growing steadily. In that case, there is a genuine independent housing fluctuation which communicates itself to the movement of aggregate income.

X. COMMERCIAL CONSTRUCTION

Commercial construction falls into two classes, construction of buildings for the use of the owners and construction for rental. There is also an intermediate class of buildings erected for rental on a long lease to a specific tenant. Buildings erected for owner occupation or for rental on a long lease to a specific tenant are generally those of a relatively specialized type, e.g., theaters, garages, department stores, and supermarkets. General-purpose office and store buildings are generally built for rental.

Investment in specialized commercial buildings ought to behave in a manner very similar to that of industrial investment, which was discussed in Chapters 4 and 5. The theory of investment in general-purpose office and store buildings is very similar to the theory of investment in apartment buildings.

A. Demand for Office and Store Space

General-purpose store space is occupied for the most part by firms in fields into which entry is very easy. There is a rapid turnover of firms in most of the trade and service fields, and there seems never to be a lack of new entrants. We can therefore suppose that the amount of space occupied depends on the number of firms which can survive under existing conditions. The number that can survive will depend primarily on the level of aggregate income and the composition of expenditure and the level of rents. We can therefore regard the demand for store space as a function of aggregate real income and rents.

Office space is occupied partly by industrial firms and partly by a variety of trade, service, and professional enterprises. It seems reasonable to suppose that the amount of space occupied by both those groups will vary with the level of income and with space rents.

Given the demand, the construction of office and store space will vary in the same way as the construction of apartment buildings. Essentially the same financial and speculative factors are at work in both cases.

NOTE ON MEASURING THE STOCK OF HOUSING

The stock of housing has many dimensions: size in terms of cubage, layout and style, equipment, location, type of construction, type of tenure, number of units per building, ground space, and condition. We cannot, therefore, readily obtain any single-figure measure of the stock of housing. Properly speaking, we ought always to talk in terms of distributions of housing characteristics. However, for our purpose it may be sufficient to adopt a much simpler view of the situation. It happens to be the case that a great many of the characteristics of houses tend to be correlated either with age or with original real cost of building adjusted for real expenditure on improvement. Generally speaking, size, equipment, quality of construction, ground space, and, to some extent, locational quality tend to be correlated with original construction cost. Style as currently valued, quality of location, quality of equipment, and condition bear an inverse correlation with age. It is obvious that there are important exceptions to this rule. There are long-period trends in style, so that older houses tend to get out of style in a systematic way; but there are also cycles, so that eighteenth-century and early-nineteenth-century houses tend to be worth more, from a style point of view, than late-nineteenth-century houses.

Neighborhoods generally deteriorate with time because of deterioration of houses, increase in traffic density, and so on, but a Beacon Hill or a Greenwich Village occasionally reverses the position. There are all sorts of random elements, depending on particular combinations of locational factors, but these elements may reasonably be taken to cancel out statistically. Of greater importance is the fact that sharp breaks in style and taste with regard to size take place. The decline in relative value of houses built in 1890 during the thirty years until 1920 was probably much greater than the decline in relative value of houses built in 1920 during the thirty years until 1950. Moreover, the relative decline was larger for the larger houses than for the smaller ones.

Nevertheless, it may be possible to get around these difficulties and construct, if not a single-figure measure, at least a one-dimensional distribution of housing stock on the basis of the correlation between the various housing dimensions and age and original cost. First, we may separate single houses from rental (multifamily housing). Second, let us ignore the problems raised by cycles and sharp breaks in style trends, and suppose that the processes of getting out of style and of neighborhood decay take place smoothly, from a statistical point of view. Under what conditions would it be appropriate to attempt to classify the housing stock in terms of depreciated value?

We can do that if it is possible to say that, from the viewpoint of its effect on the demand for new houses, a house t years old whose price-adjusted original cost was C is the equivalent of a new house costing $C(1 - pt)$ for every t and C and regardless of the composition of C. In principle, we could still get a one-dimensional distribution even if depreciation were nonlinear and if different rates applied to different levels of original cost. The difficulty then arises out of the nonhomogeneity of houses having a given age and original cost and out of the nonhomogeneity in the potential buyers from whose point of view equivalence is to be determined. From the point of view of a single individual, there is no problem if all the houses of given age and cost are essentially alike. That condition is not, strictly speaking, fulfilled, but there certainly is some tendency for all the cost factors to be correlated and for depreciation rates on the different factors to be of the same order of magnitude.

The main difficulty arises from differences in the tastes or needs of potential buyers. A large old house may be the equivalent of a new moderate-sized one to a man with a large family, and may be distinctly inferior to the new moderate-sized one to a man with a small family.

We need not, however, consider the preferences of the entire population at once. On the supposition that housing demand is not very elastic (in the long run), we can classify the population according to the value of the new house they would buy or rent (for low-income groups this involves entirely hypothetical houses, but, for reasons given below, that is not an important consideration). The economic, social, and demographic characteristics of the members of a group willing to buy or rent new houses costing a given amount will be much more homogeneous than those of the whole population. Income and family size are the major determinants of housing expenditure, and, since the range of substitution of construction quality, land cost, equipment, and size which is practically feasible is limited, the families desiring new houses of a given cost will be rather similar in characteristics (keeping single and multifamily dwellings separate). We shall not go far wrong, then, in asserting that the age–original-cost-indifference curves will be similar for all those individuals who would buy new houses of given cost (if only new houses were available).

On this basis every old house can be regarded as equivalent to some new house. However, it is possible that the same old house may turn out to be equivalent to a new house of one value from the point of view of the potential buyers of new houses of that value, and also equivalent to a new house of a different value from the point of view of potential buyers of new houses of this value. That can happen because the age–cost-indifference curves of different groups may slope differently. Cases of that sort are likely to arise only for groups whose new-house preferences are fairly similar, and therefore their quantitive importance is not likely to be so very great.

It follows that, if we may neglect errors due to (1) differences in original-cost composition of houses and difference in depreciation rates on the different cost elements, and (2) the ambiguities due to crossing of age–cost-indifference lines, every old house can be regarded as the equivalent of a new house having a certain cost, and we can describe the existing housing stock in terms of the frequency distribution of houses by new-house-equivalent cost.

It is stretching things a bit farther to aggregate such a one-dimensional distribution and to measure the aggregate supply of housing by a single figure. It is obvious that two different distributions adding to the same depreciated-value total could produce very different effects on the rate of building. Excess supply (at prices remunerative to builders) in one class will not balance excess demand in others. It will do so partially, because families will be pulled into relatively cheap houses in excess supply and out of the relatively expensive ones in short supply, but that adjustment will not be complete.

When population and income are growing, however, the situation is not so bad. The number of units demanded in nearly all classes of housing will be rising. The net change in demand for each class of house will be relatively small from year to year, and so will the net change in supply (new building aside), the change being the difference between the number leaving the class because of depreciation and the number entering by depreciation from higher classes. Within the range of classes in which building takes place, changes in the distribution of new building can fairly quickly adjust any tendency toward the appearance of oversupply in one class and undersupply in another. The main difficulty lies in dealing with the classes which consist entirely of old houses and in which new building never takes place. An oversupply of such houses may have some effect in reducing building in other classes, but this is likely to be slight because the price elasticity of housing demand is not very large. For that reason we can simply leave out of account the supply and demand for houses which are not competitive with new building.

CHAPTER 8

Personal Saving and Consumption

A discussion of personal saving and consumption runs into difficulties at the start because of the heterogeneous nature of the groups whose activities are reported statistically under the headings "personal saving" and "personal consumption." They include families whose primary source of income is wages, salaries, interest, or dividends; families whose primary source of income is an unincorporated business; farm families; and even nonprofit institutions.

The aggregate statistical data are obtained in such a way as to make it impossible to report the consumption and saving of business and nonbusiness families separately with the same accuracy as the total consumption and saving of the two together. But for analytic purposes the distinction is important and worth making even though it loosens the connection between theoretical categories and categories used in national-income statistics.

Strictly speaking, business and farm families ought to be treated separately throughout the analysis, since their behavior is governed by a mixture of business and family motives. That would be very cumbersome and, in the present state of knowledge, not very fruitful.

The easiest procedure seems to be to divide the savings of business and farm families into two parts—a strictly business part and a strictly household part. That is, we can treat family businesses and farms as though they were corporations (in their business aspect) making decisions about savings and investment without regard to the exigencies of household finance. The excess of the income of unincorporated businesses ovcr their investment in the business can be regarded as dividends, wages, and interest, which are then part of household income and subject to the same kind of consumption-savings choices as the income received by nonbusiness households.

Of course, when there is no clear separation of business and household assets, and when liquid assets are not marked as belonging to the firm or household, the measurement of business investment as distinct from household saving cannot be accurate. From the aggregate point of view, that is probably a very minor consideration.

Such a separation of business decisions from household ones is obviously too sharp. Investment in a family business is a way of providing for some of the things for which nonbusiness families save in other ways—retirement, an estate for one's heirs, and a base for borrowing against contingencies. On the other hand, the exigencies of household finance provide a more effective pressure for withdrawal of earnings from a family business than the pressures for dividends brought to bear on the management of a public corporation by stockholders or by management's recognition of stockholders' interests. One might cxpect, therefore, that the sum of household and business savings for a group of family businesses would be less than the sum of corporate savings plus personal savings from income payments by a large corporation making the same contribution to national income.

That consideration is offset by the fact that unincorporated businesses generally have lower profit rates and poorer credit resources than corporations, so that the pressure to save for business purposes is somewhat greater for family businesses than for corporations. At the same time, contractual saving in the form of life insurance and mortgage payments plays an important role in all personal saving, and business investment does not satisfy these needs. The only available evidence seems to indicate that saving in forms other than investment in their own business is about as high a percentagc of income for business owners as for others at the same income level.[1] This seems to justify the treatment suggested above.

In that case, reported personal saving consists of two parts: (1) unincorporated business saving and (2) household saving. As

[1] James N. Morgan, "The Structure of Aggregate Personal Saving," *Journal of Political Economy*, vol. 59, 1951.

Lawrence R. Klein and J. Margolis, "Statistical Studies of Unincorporated Business," *Review of Economics and Statistics*, vol. 36, February, 1954.

suggested above, we shall proceed by amalgamating unincorporated business and corporate business, making their combined profits depend on capital stocks and demand (and the other variables mentioned above). Retained earnings and (by subtraction) dividends or withdrawals for both groups are regarded as functions of profits, capital stocks, and demand. We can then discuss consumption and saving out of dividends, entrepreneurial withdrawals, and all other sources of personal income.

I. HOUSEHOLD CONSUMPTION AND SAVING

The theory of household consumption and savings decisions is very complex, and the more we learn about the behavior of households the more complex it becomes. A great number of factors appear to play a role in determining eonsumption and savings decisions of individual households: absolute current income, relative income, past income, total intangible assets, liquid assets, debts, stocks of durable goods, expectations of income and price movements, and a number of demographic variables. In formulating a "consumption function" it is necessary first of all to specify the purpose which the explanation or prediction formula is supposed to serve. If we are really interested in predicting the consumption and saving of specific individual households over short periods, all the variables mentioned will have to be taken into account. But in a prediction of aggregate consumption for a period not more than a few years ahead, the distribution of the demographic variables can be regarded as constant and need not be taken into account specifically. If the prediction applies to the aggregate over a long period, the demographic variables come into their own again.

Similarly, relative incomes may be neglected in predictions running over periods sufficiently short to permit us to regard the income distribution as approximately constant.

Assets influence saving in a number of different ways. One of the objects of saving is the accumulation of assets for retirement or for an estate. Presumably the more total salable assets a household has (other things equal) the less need it has for additional ones, and therefore its savings will tend to fall as its assets rise. The other-things-equal clause has to include the age of the family head, since perception of retirement needs and estate-building problems will change with age. It has to include also the income of the family in question and possibly the average income of the community because those variables will influence judgment on the amount of assets needed to finance retirement.

The distribution of assets by type will also be influential, particularly with respect to short-run movements of saving.[2] As long as a family wishes to spend less then the excess of income over contractual saving obligations, e.g., pension, insurance, and mortgage payments, it may not make much difference what type of assets it holds. But if a temporary decrease in income or pressure for some large expenditure makes the family consider spending more than income minus contractual obligations, the type of assets held will influence its decision. Families with a large proportion of assets invested in unincorporated businesses, housing, or insurance will be less willing to spend than families with similar total assets and income who hold their assets in the form of cash and negotiable securities. The first type of family can exploit its assets only by borrowing or selling physical assets.

Most families will refuse to sell business or housing assets except under great pressure. A considerable number of families will be less willing to take installment credit than to dispose of liquid assets, while most families will be less willing to finance consumption expenditures by mortgage debt or borrowing on insurance than by selling liquid assets. Many families do, of course, take installment credit while still holding liquid assets. They do so in part because they have greater flexibility of action when they have positive debt and positive liquid assets than when they have neither. In addition, some families feel that they are sure to pay off the installment debt, whereas they fear that, once they part with their assets, they will never be able to save enough to get them back. Even when families

[2] Morris Cohen, "Liquid Assets and the Consumption Function," *Review of Economics and Statistics*, vol. 36, 1954.

do take installment credit while holding liquid assets, their action is influenced by the presence of the liquid assets. Since the real and psychological cost of parting with liquid assets is less than that of parting with physical assets, families with liquid assets feel that they are taking less risk in taking installment credit than those with non-liquid assets.

Changes in assets may also have different effects on saving according to the way in which they occur. People who have a capital gain as a result of a change in market valuations of assets may be more inclined to increase expenditure than people whose assets have increased because they have saved. Both groups have moved toward achievement of whatever accumulation goals they may have set for themselves. Those who saved reached their present position by deliberate choice. Those who had a capital gain improved their asset position by more than they had expected and may therefore feel freer about spending a part of the increase in assets. The converse is true with respect to decreases in assets.

Changes in consumer holdings of durable goods relative to the holdings appropriate to given tastes, ages, and incomes can occur in a number of ways. An increase in income (relative to the expected increases associated with age changes) will make families feel that they can afford to have a greater annual expenditure on all types of consumers' goods, including durables. Since the durables (by definition) last a long time and have to be paid for at once, families will tend to draw down liquid assets or take installment credit, paying back themselves or the finance company out of the increased income as it accrues. This results in a sharp reduction in saving at the time of the change and in a higher saving rate at later dates than would have occurred if all increase in income were spent on nondurables.

A similar effect occurs if the development of new products leads families to reallocate their budgets as between durables and nondurables. The impact of both of these effects is considerably reduced because all families do not respond to changed circumstances at the same time.

Finally, expectations about changes in income and uncertainty about future levels of income will influence decisions on expenditures in much the same way as in the case of businesses.[3]

For short-run prediction purposes the holding of various types of assets can be measured and so, to some extent at least, can expectations. For a short period those variables can be regarded as given, and their influence can be taken into account in a prediction.[4]

When, on the other hand, we are interested in the movements of saving over a number of years, the effect of stocks of various types of assets on saving and the effect of expectations can be taken into account only if we are able to explain the movements of the asset and expectation variables themselves.

Over time, total net assets relative to income will change with changes in government debt, changes in the share of property income in total income, and changes in the valuation of equities. Of those three factors, changes in government debt are essentially exogenous, while changes in the valuation of equities are so complex that we must treat them as exogenous even though they are not. There is certainly a systematic component in the determination of the share of property incomes in total income, but the variance in the value of assets due to that cause is small relative to the variance due to the other causes. Our only course, therefore, is to treat the part of the variance in the saving ratio attributable to changes in total assets relative to income as exogenous. At least a part of the systematic element in change in the share of property income will be picked up in an equation for business saving which reflects the movements of income relative to capital stock.

The causation of movements of household holdings of liquid assets and short-term debt is even more complex. For given total assets the amount of nonliquid assets less debt held by households will depend on (1) the part of assets held in the form of houses and unincorporated business and (2) the amount of household assets held in the

[3] George Katona and E. Mueller, *Consumer Attitudes and Demand, 1950–1952,* Survey Research Center, University of Michigan, Ann Arbor, 1953.

[4] Lawrence R. Klein, "Statistical Evidence of Economic Relations from Survey Data," in Lawrence Klein (ed.), *Contribution of Survey Methods to Economics,* Columbia University Press, New York, 1954.

form of insurance equity and pension funds. For given income the first factor depends on the stocks of houses and business capital in existence and the distribution of capital between unincorporated and corporate business. The second depends on the cumulative result of decisions taken by both businesses and households over a long period in the past.

The problems resulting from an attempt to take all those complexities into account seem too great for the improvement that is likely to result. In what follows we shall attribute the variance in saving ratios due to variations in liquid-asset holdings to exogenous causes.

Variations in expectations and holdings of consumers' durables also reflect the cumulative effect of past history and a mélange of exogenous factors, and must be treated in the same way.

On the basis of those considerations we can treat personal consumption for individual families as a function of current income per head, relative income per head, lagged income, and a catchall term for the other variables which are to be treated as exogenous. It has been argued elsewhere that the effect of relative-income considerations is to make personal consumption move in proportion to personal income (for given relative-income distribution and given ratio of past per capita income to current per capita income).[5] It should be noted that if we accept Friedman's hypotheses,[6] we get the same result, though the relative-income variable plays no role.

One obtains results which are reasonably consistent with observations by using a saving function of the form $\frac{S_t}{y_t} = a\frac{y_t}{y_0} + b$, where S_t is personal saving, y_t is personal disposable income per capita, and y_0 is the highest level of personal disposable income per capita previously reached (all variables in constant prices). Equally satisfactory empirical results can be obtained by relating real personal consump-

[5] J. S. Duesenberry, *Income, Saving and the Theory of Consumer Behavior*, Harvard University Press, Cambridge, 1949.

[6] T. M. Brown, "Habit Persistence and Lags in Consumer Behavior," *Econometrica*, vol. 20, 1952.

tion per capita to lagged real personal disposable income per capita and lagged real personal consumption per capita. We can write $C_t = ay_{t-1} + bC_{t-1}$, where C_t is real personal consumption per capita and y_t is real disposable income per capita. That formulation has certain analytical advantages, and it seems reasonable to make consumption depend on a moving average of past incomes rather than on the income of a single past year.[7]

[7] L. R. Klein and A. S. Goldberger, *An Econometric Model of the United States 1929–1952,* North Holland Publishing Co., Amsterdam, 1955.

CHAPTER 9

An Income-generation Model

It was argued in Chapter 2 that the multiplier-accelerator approach to the analysis of income determination has a certain validity. That is so because of the dual relationship between investment and income. An excess of investment over Robertsonian saving (i.e., the difference between last period's income and this period's consumption) increases income. The ratio of capital stock to income is obviously a major determinant of the rate of investment. At the same time, investment increases the capital stock. One of the characteristic features of a capitalist economy, then, is a dynamic process in which the income and capital stock of one period determine the investment and consumption for the next period. The investment determines the change in capital stock from one period to the next, while investment and consumption decisions determine the change in income from one period to the next. A multiplier-accelerator process is the simplest dynamic process of that type. But though it is true that mutual adjustments between income and capital stock play an important role in the process of income determination, multiplier-accelerator systems give a deceptively oversimplified description of the real character of the interaction process. The salient features of the business-cycle theories discussed in Chapter 3 depend on the special characteristics of simple multiplier-accelerator schemes rather than on the general properties of capital-income adjustment processes.

The instability of the models proposed by Harrod and Hicks results from the use of a rigid capital coefficient and a consumption function which is independent of the returns on investment.

It was argued in Chapter 3 that a model which takes account of the relation between saving and profits and which uses a more general theory of investment would display greater stability than the multiplier-accelerator models. In the last five chapters the theories of investment and consumption sketched out in Chapter 3 have been developed more fully. It is now time to use those theories as building blocks for a theory of income determination. But before we do so, it may be useful to lay all the cards on the table at once by summarizing the conclusions of the last five chapters.

I. BUSINESS INVESTMENT

Our analysis of investment demand began with the proposition that firms behave roughly as though they equated the marginal efficiency of investment for the firm to the marginal cost of raising capital for the firm. It was argued that the schedule of the marginal efficiency of investment to the firm is the horizontal sum of two schedules. One of those schedules is the marginal-efficiency schedule for investments connected with the products which the firm is already producing. The other is the marginal-efficiency schedule for investment in producing products new to the firm. The level of the "old-product" marginal-efficiency schedule is determined by (1) the positions and shapes of the demand schedules for the firm's products, (2) the size and age distribution of the firm's stock of capital, (3) the rate of technical progress in the past, (4) the prices of capital goods relative to wage rates and the prices of other current inputs, (5) the firm's expectations with regard to future movements of demand, factor prices, and technical change. It was also argued that the old-product marginal-efficiency schedule has a considerable elasticity with respect to the rate of return.

The level of the "new-product" marginal-efficiency schedule depends in part on the same variables as the other schedule, except that the capital stocks in question are those of firms already in the industry rather than those of the firm whose investment activity is in question. In addition, however, the level of returns from investment in other lines depends on the technical and market relations between the new products and those already being produced by the firm in question. There is also the possibility that diversification will be achieved by merger rather than by real investment. Because most firms can expand in a great many lines, the new-product investment-demand schedule ought ordinarily to be very elastic. However, in a deep depression, that may not be so because diversification can be achieved more cheaply by merger than by real investment.

On the supply-of-funds side, there is reason to believe that the cost of funds is regarded by firms as low and relatively constant up to an amount equal to the sum of (1) depreciation charges, (2) "normal" retained earnings (see below), (3) a safe amount of debt which is more or less proportional to profits. The imputed marginal cost of funds rises rapidly when the firm has to take on more than the "safe" amount of debt. Ultimately the imputed cost of raising funds by debt issues equals the yield on equities. After that both equities and debt can be issued with a relatively slow rise in cost. The short-run elasticity of the supply of funds may be substantially less than the long-run elasticity, because policy on debt issues is adjusted only when the need for debt issues has been apparent for some time. Finally, firms may have a greater resistance to financing diversification by issuing debt than to financing expansion of existing product lines by that means.

For many manufacturing firms at least the rather elastic marginal-efficiency-of-investment schedule will intersect the marginal-cost-of-funds schedule in the inelastic range, but below the yield at which it pays to issue equities. In that case, the rate of investment will tend to vary with the annual rates of flow of funds from depreciation and normal retained earnings. For public utilities, however, the intersection takes place in the range of yields in which it pays to issue equities. Since the supply of funds in that range is elastic to the rate of return, the rate of investment for utilities will vary with the factors determining the marginal-efficiency-of-investment schedule (without any diversification). Some manufacturing firms may not have enough investment outlets, which yield enough to cover the opportunity cost of using retained earnings, to utilize all their normal retained earnings. They must therefore operate in the elastic range of the supply-of-funds schedule. Their rates of investment will be controlled by the factors controlling the marginal-efficiency-of-investment schedule.

To summarize a summary, the rate of investment of a firm depends on profits, retained earnings, debt in existence, depreciation charges, and on the factors determining the marginal efficiency of investment. The marginal efficiency of investment in old products depends on the relative positions of the existing cost curves for the firm's products, the new cost curves which would be obtained by buying capital, the price of capital goods, and the demand curves for the firm's products.

The position of the existing-cost curves depends on the size and age of the existing capital stock, wage rates, and prices of other inputs. The position of potential-cost curves obtainable by investing depends on wage rates, prices of other inputs, and the rate of technical change since the existing equipment was purchased. The position of the demand curves as viewed by the firm depends on past disposable income and consumer tastes, past investment decisions of other firms, and on the prices of competing and complementary products.

The marginal efficiency of investment in new products depends on the position of firms now producing the products in question in terms of the same set of variables.

Returning to the financial variables, we have argued that dividends and, therefore, retained earnings depend on past profits and past dividends. Profits themselves presumably depend on the positions of the existing cost and demand curves described above. Depreciation charges may be regarded as proportional to the existing capital stock (but when price levels change, some difficult problems of measurement arise). Debt is the cumulant of the excess of investment over depreciation and retained earnings. The stock of capital is the cumulant of investment less retirements. Finally, expectational variables are related to past events in a complicated manner.

II. INVESTMENT IN HOUSING

As we have seen, housing markets are very complicated, and their operation is not easily summarized. We can proceed most easily by successive approximations. The basic elements in the housing market are (1) the demand for houses or apartments, (2) the existing supply of houses, (3) the supply curve for new construction. The position of the price-quantity-demand schedule for houses is primarily dependent on the level of income per family and on the number of families. There is some, but not very much, price elasticity of demand. The supply of houses in existence is primarily the result of past construction and demolition. The amount of housing which would go off the market for any measurable decline in price is probably very small. The rate of new construction depends on the level of prices or rents and on the level of construction costs. In a very simple model of the housing market, then, the price level of houses (and the level of rents) is set in such a way as to equate the demand for existing houses to the supply of existing houses. The level of construction costs determines the position of the supply curve of new building. The rate of building is, then, such as to equate the supply price of building with the going price for houses determined in the market for existing housing. It is easy to see, then, that in such a market the rate of building tends to increase with disposable income and population, to decrease with the size of the existing stock, and to decrease as the level of construction costs rises.

The changes in construction cost are exogenous. No doubt the rate of building does have an effect on the cost of construction, but the connection is very complicated and must be ignored here.

At a first approximation, then, the rate of investment in housing is a function of disposable income and the depreciated stock of housing, with population playing an important role as an exogenous variable.

But, as we have seen, the real situation is very much more complicated because of the lags in the system. Three sets of lags have to be taken into account: (1) lagged adjustments of prices and rents to the supply-and-demand situation, (2) lags in the entry of new firms into the building industry, (3) lags in the response of demand to changes in income because of the "down-payment lag."

In addition, it is necessary to take account of the influence of speculative activity in the construction of apartments. Very similar considerations apply to the construction of office and store space.

III. CONSUMPTION

Our summary of consumption theory can be very brief. In Chapter 8 it was argued that the rate of consumer expenditures is determined largely by disposable income (with a short lag) and past consumption. A great many other variables influence consumption, but they are nearly all exogenous (or at least have to be treated as exogenous). Disposable income has to be defined so that it does not include the income of entrepreneurs reinvested in their own businesses. In other words, entrepreneurial withdrawals are treated as dividends.

IV. ENDOGENOUS AND EXOGENOUS VARIABLES

Our object is to use the generalizations made in the last three sections to show how the movements of income can be explained. To do that it will be necessary to simplify the theory of investment and consumption and to organize our analysis in some convenient way. A large number of variables were introduced in the discussion of investment and consumption. We shall divide those variables into two classes: (1) A set of variables, each of which is strongly influenced by the current or lagged values of other variables in the set (or by its own lagged values). The variables in this set are called "endogenous variables." The set of interrelations among them is called the "endogenous system." (2) The remaining variables, which are regarded as having an influence on the endogenous variables but as not being much influenced by the endogenous variables. The members of this second set of variables are called "exogenous variables."

A. Choice of Exogenous Variables

The decision to regard a variable as exogenous may be taken on any of a number of grounds. In some cases the decision is necessary. We simply do not know how to explain the variable. We must take its value at any given time as arbitrary. A second group of exogenous variables consists of policy variables. We might be able to give some explanation of the variable's movements, but we are interested in studying the influence of the variable on the rest of the system. Even if there were a very precise connection between Federal Reserve policy and the rate of change of prices, we would regard monetary policy as exogenous because we want to know how things would work out under a different regime.

In a third class of cases we consider a variable to be exogenous because its explanation would lead us too far afield. Population growth may be explained in some way, but in order to explain it, one would have to introduce a great many variables and complicate the income-determination model enormously. It seems more convenient to take population as a given datum and then explain its movement separately.

A very closely related group of cases involves unique events. An event like the railroad boom of the 1870s is a case in point. Given the institutional arrangements of the United States after the Civil War, the rate of population growth, the availability of free land, and the economic characteristics of railroads, the boom was in a certain sense inevitable. But there will never be another one just like it or even approximately like it. The same thing could be said of the stock-market boom of the 1920s and of many other events. A system which contained enough variables and relationships to explain events of that type would be too cumbersome to yield any understandable results.

On the other side of the coin, there are certain variables which ought not to be treated as exogenous. If we are trying to explain income and believe that the rate of investment is strongly influenced by the level of income, it does not make much sense to treat investment as an exogenous variable. In general, we ought to try to include in our endogenous system any variables which interact strongly and systematically with other variables already in the system. There are, therefore, certain general criteria for deciding whether to treat a variable as endogenous or exogenous, but there is also an arbitrary element of judgment and convenience in the choice.

When one variable, say, x_1, is a function of another variable, x_2, and x_2 is exogenous, we often find it convenient to treat x_1 as exogenous.

B. Parameters as Exogenous Variables

Exogenous variables sometimes appear simply as constant terms in the relationships determining the endogenous variables. For example, government purchases of goods and services is a constant term in the definitional equation for gross national product. But the relations among the endogenous variables also contain parameters, e.g., the slope of the consumption function. Those parameters are not fixed over time like the constants of physics. They are themselves the outcome of the whole process of social evolution. Unless we produce a general sociology, we must take the outcome of that process as a given datum. For purposes of analysis we regard the parameters in our behavior relations as fixed, but we must keep in mind that they are, in fact, changing over time. Some implications of that notion are discussed in the next chapter.

V. INTERINDUSTRY RELATIONS

Our discussion of the theory of business investment applies to the individual firm or industry. To use the theory in an income-generation model we must now consider the interaction among firms. At the same time we shall classify the determinants of investment into endogenous and exogenous variables. This classification will help to simplify the interfirm connections.

Each firm's investment decisions are conditioned in a number of ways by the prices being charged by other firms. The price being charged today by any one firm presumably depends on the position of its cost curve (which in turn depends on its own stocks of capital and prices charged by others a short time ago), on its recent sales, and on prices recently charged by rivals.

Each firm is linked to other firms in the system in a number of complex ways: (1) Its sales are generated partly by the investment orders of others; (2) its sales are influenced by the input coefficients of others, which depend on the techniques being used; (3) its sales and its view of the elasticity of its demand depend on prices charged by others; (4) its costs depend on prices charged by suppliers; (5) the price of capital goods which it might buy is set by other firms; (6) return from investments in new products depends on conditions in other sectors. The price variables in the system are mostly in relative form. A general proportional increase in all wages and prices in the system decreases the burden of debt and, because of lags in the adjustment of tax rates and depreciation charges, has some non-proportional effect on reported profits and real profits after tax. Many important and interesting questions arise in this connection, but they are not considered in this book.

Relative price changes arise in several different ways: (1) through differences in rates of change of technique between industries; (2) through changes in competitive structure; (3) through changes in tax and subsidy arrangements; (4) through differential changes in labor's bargaining power due to changes in the position of unions; (5) through differential changes in wages due to differences in labor-market conditions faced by different industries, e.g., differences in the level of employment in the areas in which they are located; (6) through changes in the utilization of capacity.

The first four causes of relative price changes are essentially exogenous. Since they are not explained by any other elements in the system, they can be treated as members of the large class of causes of shifts in the relations among the endogenously determined variables of the economic system. The fifth cause is also basically exogenous. Differences in natural increase of population in different areas are exogenous, and so, for the most part, are differences in rates of growth of sales and thereby of demand for labor. Differences in utilization rates may be due to exogenous causes, but the capacity-utilization rates depend in part on the size of the capital stock in the industry. Since the capital stocks result from investment which we are trying to explain, we cannot say that they are exogenously determined. But the relative price changes arising from changes in relative utilization rates are likely to be small. The effect of price changes induced by changes in capacity utilization on profits may be substantial, but the effect of the same price change on the positions of the demand and cost curves of other firms is small enough that we can neglect it.

In what follows, therefore, we shall implicitly take account of price changes due to changes in capacity utilization in our explanation of profits. But we shall ignore the secondary repercussions of those price changes on the distribution of demand among firms. We may also ignore the effect of changes in one firm's price (when due to a change in the utilization rate) on the costs and thereby on the investment behavior of other firms. Finally, we ignore the effect of changes in capacity utilization on the prices of investment goods relative to wages.

If technique is improving over time, it is almost certain that either prices will fall or money wages will rise. For simplicity, we shall assume that wages generally tend to rise with productivity, so that there is no systematic tendency for prices to fall because of secular trends in technique. That does not mean that labor's share of gross or net national product must always be the same. It can vary if the operation of a given endogenous model causes variations in the ratio of capital stocks to demand. Labor's share can also vary because of changes in the structure of the model. What we have ruled out by the constant-price-level assumption is the possibility that cost reductions do not result in price cuts even when capacity utilization is constant. Such a state of affairs is certainly possible and may even be important over short periods. It is ruled out here for reasons of simplicity and a somewhat unsupported judgment that it is not important.

On the strength of those rationalizations, we do not need to consider prices explicitly in an endogenous model. The distributional effects of price changes due to changes in utilization of capacity are implicitly taken into account in the profits function.

VI. OTHER INTERINDUSTRY CONNECTIONS

The firms in the economy are linked to one another through the input-output connection, i.e., through the fact that some firms supply others with materials. In principle, then, the sales of any one industry or firm are dependent upon the previous sales of all other industries or firms. That connection can be simplified by making each industry's sales equal to the sum of the sales of each final product multiplied by the appropriate coefficient of the inverse of the input-output matrix. The final-product sales must, of course, be lagged to allow for the lapse of time between sales of final product and sales of "intermediate" products used to produce them.

The effect of this simplification is to eliminate the whole problem of the relations between sales, production plans, inventories, and the placement of orders for supplies. In a proper model, we would have to make each firm's production depend on the volume of orders on hand, the rate of receipt of orders from other firms over some recent period, and the existing inventory of finished products and work in process. The same variables would also determine production plans for the future. Those plans plus the current inventory of raw materials and the volume of outstanding orders for materials would determine the rate of placement of new orders. There would be a time lag between changes in the rate of sale of final products and changes in production, because orders are placed and production schedules changed as finite intervals. The amount of change in production resulting from a change in the rate of sale of final products will depend on the initial situation in terms of inventories and orders on hand for the firms in the system. The relation between rates of production and rates of sale of final products will be the same as the one calculated by inverting an input-output matrix if the rate of final-product sale remains constant for a sufficient length of time to permit complete adjustment of inventories and orders outstanding.

If we make each firm's sales proportional to lagged final-product sales (counting planned additions to inventory as part of final-product sales), we suppress all the intricate linkages mentioned above. In addition, we eliminate the possibility of giving systematic consideration to the effects of changes in expectations on the relations between sales and production. That gap is filled in a somewhat *ad hoc* way in the discussion of income fluctuations in Chapter 11.

Even if we make each firm's sales directly dependent on final-product sales, we still find that each firm is directly linked to every other firm. The final-product sales which generate demand for any one firm's product include sales of capital goods to other firms. Since different firms buy different types of investment goods, the sales of any one firm today depend on the sales, capital stock, and other investment-determining variables for other firms at some time in the past. We shall bypass this set of interfirm linkages by making each firm's sales simply proportional to gross national product.

That means that we must neglect the effect of cyclical changes in the ratio of investment to GNP on the distribution of sales among industries. The significance of the interindustry distribution of sales lies in the fact that a gain or loss of sales in one industry has a greater effect on investment than a similar gain or loss in other industries. No substantial error is introduced by linking sales to GNP unless the industries which supply capital goods are substantially more capital-intensive than those which supply consumers' goods. There seems to be no evidence that that is the case.

Of course, the sales of different types of consumers' goods do not vary in proportion to GNP either. The argument just given applies to differences in cyclical sensitivity of demand for consumers' goods except for housing. Housing is so much more capital-intensive than other consumers' goods that it requires special treatment. A discussion of the longer-run effects of differences in the income elasticities of demand for consumers' goods is given in Chapter 10. Pending that discussion, it will be assumed that each firm's sales are proportional to GNP.

The last interfirm connection mentioned above is the link between conditions in one industry and opportunities for diversification in another. It is also important because it reduces the effect of differences in rates of growth of demand on the profits of particular industries, and reduces the chance that some industries will be unable to invest for lack of funds, while others are unable to find profitable outlets for the funds they have. A disaggregated model which ignores the "diversification linkage" is in error because it exaggerates the extent to which profits earned in one industry are plowed back into the same industry. That does not raise very serious problems here. In what follows we shall use aggregative models extensively, modifying the results by reference to some of the properties of a disaggregated model.

In short, we have suppressed all six of the interindustry linkages listed above. As a result, our investment theory makes each firm's investment depend on (1) GNP, (2) its own capital stock and its age distribution, (3) profits and/or retained earnings, (4) outstanding debt, (5) past rates of technical change, (6) expectational variables. The last two variables are treated as exogenous. The age distribution of the capital stock can be roughly taken into account by using depreciated rather than gross capital stock. Some of the problems raised by Domar in this connection are discussed below.

For fixed values of the rate of technical progress and expectational variables, investment by a firm can be regarded as a function of lagged GNP, lagged depreciated capital stock, lagged profits or retained earnings, and outstanding debt. Other things equal, profits are a function of GNP and capital stock. Retained earnings are dependent on lagged profits and lagged dividends. The parameters of the functional relationships are, of course, different for different firms and are subject to change for a variety of reasons already mentioned. Later we shall use those functional relationships in a larger system of equations to determine the movements of aggregate income.

In discussing consumption we concluded that all influences on consumption except disposable income and past consumption should be treated as exogenous.

In dealing with housing investment we shall take population as an exogenous variable. At the risk of some error, we shall take the level of construction costs as exogenous as well. The rate of housing construction is therefore determined by the level of disposable income, the size of the existing stock of houses, the average level of income over some time in the past, and the lagged rate of construction.

VII. AN ENDOGENOUS SYSTEM

The propositions made in the last few pages can be translated into mathematical form. The resulting equations can be put together to form an endogenous model of the income-determination process. The notation used is as follows:

I_B—business investment
I_H—housing investment
C—consumption expenditure
Y—gross national product
R—capital-consumption allowances
K_B—business capital stock (net of depreciation)
K_H—stock of houses (net of depreciation)
π—profits (including those of unincorporated businesses and farms)
d—dividends and entrepreneurial withdrawals
E—retained earnings of business (corporate and unincorporated)
D—business debt
Y_d—personal disposable income

The argument presented above leads to the following set of equations:

$$I_{Bt} = f(Y_{t-1}, K_{Bt-1}, E_{t-1}, \pi_{t-1}, D_{t-1}, R_{t-1}) \tag{1}$$

which represents the dependence of business investment on income and capital stock as determinants of the marginal efficiency of investment and on retained earnings, profits, depreciation, and debt as determinants of the cost of raising funds.

$$I_{Ht} = f(Y_{t-1}, K_{Ht-1}) \tag{2}$$

which represents the dependence of the rate of housing construction on the age of the existing stock of houses. Population affects the constants in the equation. The longer lags have been left out for the moment.

$$C_t = \phi(Y_{dt-1}, C_{t-1}) \tag{3}$$

This represents the dependence of consumption on disposable income and lagged consumption.

(4) $$\pi_t = \theta(Y_t, K_{Bt-1})$$

This indicates that profits (in dollars per year) depend on capital stock and aggregate demand.

(5) $$d_t = \chi(\pi_{t-1}, d_{t-1})$$

Dividends depend on lagged profits and lagged dividends.

The remaining equations are definitional.

(6) $$Y_t = I_{Bt} + I_{Ht} + C_t$$

i.e., GNP is the sum of investment and consumption.

(7) $$Y_{dt} = Y_t - \pi_t + d_t - R_{Bt} - R_{Ht}$$

i.e., disposable income equals GNP less net profits and depreciation allowances, plus dividends and entrepreneurial withdrawals.

(8) $$R_{Bt} = k_B K_{Bt}$$

Business depreciation is proportional to business capital stock.

(9) $$R_{Ht} = k_H K_{Ht}$$

Housing depreciation is proportional to housing stock.

(10) $$K_{Bt} = K_{Bt-1} + I_{Ht} - R_{Ht}$$

(11) $$K_{Ht} = K_{Ht-1} + I_{Ht} - R_{Ht}$$

Capital stock at the end of period t equals capital stock at the end of period $t - 1$ plus investment during period t less depreciation during period t.

(12) $$E_t = \pi_t - d_t$$

Retained earnings equal profits less dividends.

(13) $$D_t = D_{t-1} + I_{Bt} - R_t - E_t$$

Business debt at the end of period t equals business debt at the end of period $t - 1$ plus investment during period t less depreciation and retained earnings during period t.

This system of equations describes a dynamic process. If we specify values of the twelve variables in the equations at any one time, we can calculate the values for the next period, and so on. Thus, the equations (if numerically specified) and the initial conditions determine the path of movement of income through time which would result if the values of the exogenous variables remained fixed. Of course, the actual path of movement will depend on the movements of the exogenous variables as well as on the properties of the equations.

In the next three chapters we shall discuss the interaction of the exogenous variables with the dynamic process generated by a given set of equations. To make the argument intuitively understandable and to minimize the use of mathematics, we shall use a simplified version of the model given above and then reintroduce the complications one at a time.

The simplified model is derived from the one above in the following way:

1. Eliminate the housing equations so that we have to deal with only one type of capital.

2. Simplify the lag structure by making consumption depend on disposable income alone, leaving out lagged consumption. Similarly, make dividends dependent on profits alone, leaving out lagged dividends.

3. Suppose that consumption is influenced by current rather than lagged dividends (since dividends change slowly, that has little effect).

4. Drop the debt variable out of the investment equation.

5. Drop the retained-earnings variable out of the investment equation so that investment depends on profits but not on retained earnings.

6. Make profits depend on the capital stock at the end of the period rather than at the beginning. Since capital stock changes very slowly, this has a relatively small effect.

Those simplifications cause the following changes in the equations:

1. Equations (2), (9), and (11) are dropped, and terms R_H, I_H are dropped from equations (6) and (7) by the elimination of housing.

2. Equations (12) and (13) drop out when we leave out debt and retained earnings in the investment equation. The variables D_{-1} and E_{-1} are eliminated from equation (1).

3. The terms C_{-1} and d_{-1} disappear from the equation for consumption (3) and the dividend equation (5).

4. The profits equation (4) is rewritten as

$$\pi_t = \theta(Y_t, K_{Bt}) \tag{4a}$$

5. The consumption equation is rewritten by substituting from (5) and (7).

$$C_t = \Phi(Y_{t-1} - \pi_{t-1} - R_{Bt-1} + d_t)$$

We now have the following seven equations:

$$\begin{array}{ll}
(1) & I_t = f(Y_{t-1}, K_{t-1}, \pi_{t-1}, R_t) \\
(2) & C_t = \Phi(Y_{t-1} - \pi_{t-1} - R_{t-1} + d_t) \\
(3) & d_t = \chi(\pi_{t-1}) \\
(4) & \pi_t = \theta(Y_t, K_t) \\
(5) & R_t = kK_{t-1} \\
(6) & K_t = K_{t-1} + I_t - R_t \\
(7) & Y_t = I_t + C_t
\end{array}$$

(Since there is only one type of capital and one type of investment, the B subscripts have been dropped from K, I, and R.)

Those equations can be further reduced by substitution. The profits variable can be eliminated by substituting from equation (4) into equations (1) and (2). The dividend variable can be eliminated by substituting from equation (3) in equation (2). The depreciation variable can be eliminated by substituting from equation (5) in equation (1), (2), and (6).

Those substitutions reduce the system to the following four equations:

$$\begin{array}{ll}
(1a) & I_t = f[Y_{t-1}, K_{t-1}, \theta(Y_{t-1}, K_{t-1}), kK_{t-1}] \\
(2a) & C_t = \Phi[(Y_{t-1} - \theta(Y_{t-1}, K_{t-1}) - kK_{t-1}), \chi(\theta(Y_{t-1}, K_{t-1}))] \\
(3a) & K_t = K_{t-1} + I_t - kK_{t-1} \\
(4a) & Y_t = I_t + C_t
\end{array}$$

Let us now, for simplicity, represent these equations in linear homogeneous form. Then, since investment depends only on K_{t-1} and Y_{t-1}, we have

$$I_t = \alpha Y_{t-1} + \beta K_{t-1} \tag{1b}$$
$$C_t = aY_{t-1} + bK_{t-1} \tag{2b}$$
$$K_t = (1 - k)K_{t-1} + I_t \tag{3b}$$
$$Y_t = I_t + C_t \tag{4b}$$

If we now substitute equations (1*b*) and (2*b*) in (4*b*), we obtain

$$Y_t = (\alpha + a)Y_{t-1} + (\beta + b)K_{t-1} \tag{4c}$$

Similarly, by substituting (1*b*) in (3*b*), we obtain

$$K_t = \alpha Y_{t-1} + [\beta + (1 - k)]K_{t-1} \tag{3c}$$

We shall make use of these equations in the next chapter in working out the dynamics of income determination. Meanwhile, it is worth noting a few points about the significance of the parameters. Each of the four parameters α, β, a, b reflects the influence of a number of different factors. a is the marginal propensity to consume out of GNP. An increase in the marginal propensity to consume out of disposable income increases a. But a also reflects the increase in profits which results (*cet. par.*) from an increase in income. An increase in the sensitivity of profits to changes in income therefore reduces a. That influence is partly offset by the effect of profits on dividends and the effect of changes in dividends on consumption.

b reflects the influence of changes in capital stock on consumption. That influence results from the influence of capital stock on profits through the influence of profits on dividends and dividends on consumption.

α represents the effect of changes in income on investment, which is felt in two ways. A change in income changes the marginal efficiency of investment, and it also changes the marginal cost of funds by changing profits. β reflects the influence of capital stock on investment working through both marginal efficiency and profits.

From the last two equations we can derive a recursive equation relating income at time t to income in the two preceding periods.[1]

$$Y_t = [(\alpha + a) + (\beta + 1 - k)]Y_{t-1} + [\alpha(\beta + b) - (\alpha + a)(\beta + 1 - k)]Y_{t-2}$$

This is a generalized version of a multiplier-accelerator process. In the simple multiplier-accelerator process the investment equation is written as $I_t = \alpha Y_{t-1} - K_{t-1}$, so that $\beta = -1$. Depreciation, is usually neglected and investment treated as net investment, so that $k = 0$. The connection between the capital stock and consumption is neglected, so that $b = 0$. If we substitute those values in the equation for Y_t, $(\beta + 1 - k) = 0$, $\alpha(\beta + b) = 0$,

$$Y_t = (\alpha + a)Y_{t-1} + \alpha(-1)Y_{t-2}$$

the usual form of the equation for the multiplier-accelerator process.

The a in our equation will be much smaller than the marginal propensity to consume out of disposable income because it reflects the influence of income changes on profits and business savings. At the same time the α in our equation will be much less than the average ratio of capital to output (which is the accelerator in multiplier-accelerator models). A \$1 increase in income (capital-stock constant) will increase the rate of business investment by an amount which is not much larger than the increase in business savings resulting from a \$1 increase in income. This amount will surely be only of the order of magnitude of, say, \$0.25. Even when we

[1] The proof is as follows: From (4c) and (3c) we have

(1) $$Y_t = (\alpha + a)Y_{t-1} + (\beta + b)K_{t-1}$$
(2) $$K_t = \alpha Y_{t-1} + (\beta + 1 - k)K_{t-1}$$

Multiplying (1) by $(\beta + 1 - k)$ and (2) by $(\beta + b)$ we obtain:

(1a) $$(\beta + 1 - k)Y_t = (\alpha + a)(\beta + 1 - k)Y_{t-1} + (\beta + b)(\beta + 1 - k)K_{t-1}$$
(2a) $$(\beta + b)K_t = \alpha(\beta + b)Y_{t-1} + (\beta + b)(\beta + 1 - k)K_{t-1}$$

Subtracting (2a) from (1a);

(3) $$(\beta + 1 - k)Y_t - (\beta + b)K_t = [(\alpha + a)(\beta + 1 - k) - \alpha(\beta + b)]Y_{t-1}$$
(3a) $$(\beta + b)K_t = (\beta + 1 - k)Y_t + [\alpha(\beta + b) - (\alpha + a)(\beta + 1 - k)]Y_{t-1}$$

Substituting from (3a) in (1),

$$Y_t = (\alpha + a) + (\beta + 1 - k)Y_{t-1} + [\alpha(\beta + b) - (\alpha + a)(\beta + 1 - k)]Y_{t-2}$$

allow for the effect of the increase in income on housing construction, the total value of α will surely be less than the average capital coefficient. As a consequence of those considerations, an increase in income will have a much smaller immediate effect on expenditure than would occur in a simple multiplier-accelerator model. That conclusion will be reinforced when we take the lags in dividend payments, construction, and housing demand into account.

On the other side of the coin, the negative effect of an increase in capital stock (income constant) will be much smaller than the one assumed in multiplier-accelerator models. A $1 increase in business capital stock (income constant) will reduce profits by a very small amount and will have a correspondingly small effect on business investment. Part of the decline in business investment will be offset by a decline in business saving.

The effect of those considerations is to reduce the immediate effect of an increase in income on expenditure but to make the effect persist for some time, since investment will decline slowly (if there is no further increase in income) as capital accumulates. The system will therefore be much more stable than a simple multiplier-accelerator system.

The implications of those considerations are developed in detail in the next three chapters.

CHAPTER 10

The Equilibrium Rate of Growth

A closed model like the one outlined at the end of the last chapter has a limited usefulness. From the way in which the model was constructed it is apparent that the model cannot in itself explain the actual movements of income. The model provides the basis for analyzing the interactions among a set of variables which influence one another to a marked extent. But many of the variables in the system are also influenced by other variables whose movements are not readily explained by the methods in which economists specialize. For any given values of those exogenous variables the closed or endogenous system of equations of the last chapter will have a fixed set of numerical parameters. With given parameters and given initial values for the variables, the system of equations will generate a sequence of values for the variables in the system. But since the exogenous variables do not in fact remain constant, the sequence generated by any one set of numerically specified equations will never materialize. The reader will recall that the exogenous variables here are not restricted to such items as population, government expenditure, and taxes, which appear in econometric models as exogenous variables. They also include variables such as the form of business organization and the structure of financial institutions which affect the parameters of the behavior relations in an endogenous model. It may be taken for granted that those variables are constantly changing.

In view of the constant changes in the exogenous variables, we ought to suppose that any one numerically specified model of the economy holds for only a short period. In a proper analysis of economic history we would proceed in the following manner if we could: (1) At any point in time the exogenous variables have certain "values" which determine the parameters of the endogenous equations. (2) At the same moment the variables in the endogenous equations have certain initial values inherited from the past. (3) The parameters and initial values of the variables in the endogenous system determine the movement of the endogenous variables over a short period. (4) That movement determines the initial conditions for the next period. (5) Meanwhile the exogenous variables will change and change the parameters of the system. An account ot the history of the variables in the endogenous system would consist of a series of statements about the movements of the exogenous variables and the consequential movements of the endogenous variables.

Such a procedure would be ideal but obviously would be impossible to follow. We must seek an approach which requires less information and less detailed computation but which still reflects the occurrence of changes in the economic structure. To begin with, we can divide parameter changes into three classes: (1) Many parameters have a fairly well marked trend. (2) There are random deviations from the trend values or from the constant means of those parameters which do not have a trend. (3) There are rather rapid movements in parameters as a result of the development and collapse of speculative booms, monetary crises, and the like. Fortunately movements of the last type are relatively infrequent.

If we can neglect the random changes, we can divide our analysis into two parts. We first consider a world in which parameter changes take place smoothly. We can then show how the rapid shifts in parameters (which will be called "shocks") influence the system.

In dealing with smooth changes in parameters we can use what might be called the "order-of-magnitude approach." To indicate the nature of the approach, let us first consider a world which can be described in terms of a model like the one outlined in Chapter 9 but on a disaggregated basis with separate investment and business-saving equations for each sector. The system will then have to contain equations for the distribution of demand among sectors, and population will appear as a parameter in some of the equations. Suppose that the parameters of the model are changing over time, but that the changes occur smoothly and with no changes in the direction of change.

The fact that the parameters are changing influences the behavior of the system in two different ways.

1. The occurrence of changes in the parameters will gradually change the characteristic behavior of the system. Suppose, for example, that we had a simple multiplier-accelerator system with a slowly rising propensity to save. At one date the system might tend to produce exponential growth in income. As time passed, the increase in the propensity to save would change the character of the system so that eventually it would no longer be capable of exponential growth. It would begin to produce antidamped cycles. Under appropriate circumstances that difficulty can be met in the following way. If the parameter changes take place slowly, it may be possible to divide history into rather arbitrary periods. We then say that in a certain period the parameters of the economic system had a certain order of magnitude. We can then consider the characteristic behavior of systems with parameters having the order of magnitude in question. The description of behavior would, of course, have to run in order-of-magnitude terms. That is, we should say that a system with parameters lying in a certain range tends to produce exponential growth in income at a rate of growth lying in a certain range. We would then analyze history by showing, first, that income movements in a certain period result from the characteristics of the economic system and, second, that movements of a different type took place in a later period because the parameters of the economic structure had changed.

A complete analysis of economic history would require an explanation of the parameter changes themselves. Such an explanation is at least as important as an elucidation of the consequences of the existence of a given set of behavior parameters, but it requires talents of a different type from those required to analyze models with given parameters.

2. It was said above that the occurrence of changes in parameters affects the behavior of the system in two ways. As we have seen, the fact that the levels of certain parameters are changing will

change the behavior of the system as time passes. Over a relatively short period we may take the levels of the parameters to be fixed. But a system with parameters fixed at a given level may not behave in the same way even over a short period as a system with parameters which are at the same level but changing. As we shall see, a system with a rising population is significantly different from a system with a constant population of the same size.

So long as parameter changes are smooth, we can overcome that difficulty by expanding our model so that the rates of change of some parameters become parameters of an expanded system. (Technically this device enables us to convert a nonhomogeneous system into a homogeneous one.)

If there were no shocks, then, and if all parameter changes took place smoothly, we could explain the movements of income by reference to the characteristics of a sequence of economic systems each having parameters of a certain order of magnitude. To do that we should at least have to know the order of magnitude of the parameters in each time period.

Moreover, since there are shocks, i.e., sudden changes in parameters, we should have to identify them concretely and show how they affected the operation of the system. In fact, we do not have adequate estimates of the parameters obtaining now, let alone estimates for those obtaining in past periods. We must make a roundabout attack on the explanation of observed movements of income. We can proceed in the following way. We begin by supposing that the economic system at any time since the Civil War can be described qualitatively in terms of the model developed in Chapter 9 (with certain additions to allow for population growth and change in distribution of demand among sectors). The use of that model is justified by our earlier discussion of investment and consumption behavior. We can use that model to find out what kinds of income movements can occur and how various changes in the economic structure affect the movements of income. We know what kinds of income movement have occurred. We also know something about the changes in the economy which have occurred. If the model is qualitatively correct, its general characteristics plus the observed changes in the structure of the system should imply changes in the type of income movement which have occurred. If the model passes that test, it will contribute something to our understanding of what may happen in the future.

In this chapter we shall assume that all parameter changes take place smoothly, and shall confine our attention to the analysis of the effects of structural changes on the rate of growth of income. In the next chapter we consider shocks and try to show how the structure of the system determines the system's response to shocks.

In Chapter 12 we try to show (1) that shocks resulting from shortages of labor, capital, or credit are not in themselves likely to produce major depressions; (2) that the depressions before 1929 were caused primarily by the collapse of speculative booms, monetary crises, and similar factors; (3) that by the 1920s the structure of the American economy had changed in such a way that income tended to grow more slowly than it did at the beginning of the century. At the same time the changes in the structure of the system reduced its stability so that adverse shocks had a greater effect than formerly. Various special factors helped to maintain the rate of growth of income during the 1920s, but they made the system less stable than it otherwise would have been. The proximate cause of the depression of the 1930s was the collapse of a speculative boom, but a depression might have occurred anyway.

The general nature of the argument leading up to those conclusions can be indicated by the simple model developed at the end of Chapter 9.

I. A SIMPLE DYNAMIC MODEL

At the end of Chapter 9 we worked out a very simple generalized multiplier-accelerator model. The final equations of the model were:

$$y_t = (\alpha + a)y_{t-1} + (\beta + b)k_{t-1} \tag{1}$$

$$k_t = \alpha y_{t-1} + (\beta + 1 - l)_{t-1} \tag{2}$$

It is obvious that a model of this sort will generate a sequence of values for income and capital stock from any set of initial values for those variables. If the values of k and y are given for period zero, they can be substituted into the equations as values for k_{t-1} and y_{t-1}. The equations can then be used to calculate the values of income and stock of capital in period 1, and the process repeated indefinitely. We want to find what kinds of sequences can be generated by a model of this type and how the properties of the sequences are related to the parameters of the system and to the initial conditions.

Because the model is so simple, it can generate only very simple sequences. We can see fairly readily what the possibilities are by examining the relative growth rates for capital and income. The capital stock grows whenever gross investment exceeds depreciation. Income grows when investment exceeds Robertsonian saving. In terms of our equations the rate of growth of income is given by

$$r_y = \frac{y_t - y_{t-1}}{y_{t-1}} = (\alpha + a - 1) + (\beta + b)\frac{k_{t-1}}{y_{t-1}}$$

where r_y is the rate of growth of income, or $\frac{y_t - y_{t-1}}{y_{t-1}}$. The rate of growth of capital stock is given by

$$r_k = \frac{k_t - k_{t-1}}{k_{t-1}} = \frac{\alpha y_{t-1}}{k_{t-1}} + (\beta - l)$$

where r_k, the rate of growth of capital, is $\frac{k_t - k_{t-1}}{k_{t-1}}$. It will be observed that the two rates of growth are determined entirely by the ratio of capital stock to income. That being the case, the two rates of growth will remain constant if the ratio $\frac{k_t}{y_t}$ remains constant. The ratio of capital to income will remain constant if the rate of growth of capital equals the rate of growth of income. Consequently income *can* grow steadily if there is a real, positive ratio of capital to income at which the rate of growth of income equals the rate of growth of capital.

It must be kept in mind here that the ratio of capital to income is not technically determined. It is simply the result of the cumulative effects of investment on the capital stock and of investment and consumption on income. As we pass from low to high ratios of capital to income the rate of growth of capital tends to fall (*a*) because the returns on investment are reduced by an increase in the capital-output ratio (with a given state of technique), (*b*) because profits per unit of capital tend to fall, (*c*) because a given ratio of investment to income produces a smaller percentage increase in capital stock. At the same time as we pass from low to high capital-income ratios, the rate of growth of income tends to fall. For the reasons just given, the ratio of investment to income tends to fall as the capital-income ratio rises. However, the ratio of saving to (lagged) income also tends to fall because of the decline in the ratio of profits to income. In general, the first effect tends to outweigh the second. That is so because investment is almost certain to fall by more than business saving when the capital-income ratio rises. (That is so at least in terms of a system with no "autonomous" investment and no dividend lag. Exceptions may occur when those factors are taken into account. They are discussed below. There may also be exceptions when we consider a multisector system.)

We can assume, then, that the rate of growth of capital and the rate of growth of income decline as the ratio of capital to income rises. Income can grow steadily if there are points at which the two rates of growth are equal to one another. Any values of the capital-output ratio at which the rate of growth of income equals the rate of growth of capital may be called "equilibrium points." The term "equilibrium" simply means a state of affairs which maintains itself once established. Correspondingly, the rates of growth of capital and income at the equilibrium values of the capital-output ratios can be called "equilibrium rates of growth."

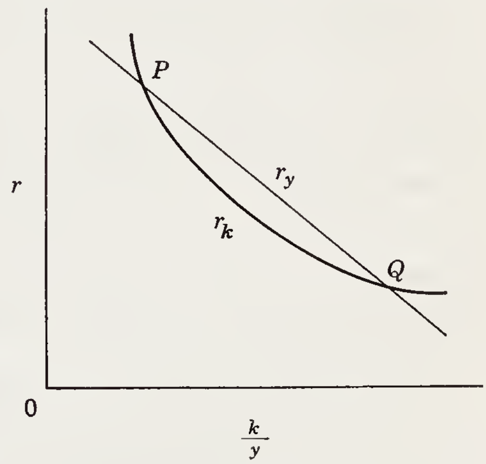

FIGURE 9.

The interrelation between the rates of growth of income and capital and the ratio of capital to income are illustrated in Figure 9.

From the general properties of the equations for the rates of growth of income and capital we know that r_k (rate of growth of capital) is a hyperbolic function of $\frac{k}{y}$, while r_y (rate of growth of income) is a linear function of $\frac{k}{y}$. It follows the r_k curve intersects r_y either twice or not at all (except in the special case of tangency of the two functions). Of the two intersections, one corresponds to a stable equilibrium position, the other to an unstable one. This can be seen by reference to Figure 9. The intersection of r_k and r_y at P is stable. If $\frac{k}{y}$ lies to the left of P, r_k exceeds r_y. k therefore grows faster than y, and $\frac{k}{y}$ rises until it reaches P. Similarly, if $\frac{k}{y}$ lies between P and Q, r_y exceeds r_k, and the ratio of k to y will fall until P is reached. If $\frac{k}{y}$ lies to the right of Q, r_k again exceeds r_y, and $\frac{k}{y}$ rises. r_y and r_k fall until r_k attains the maximum negative value permitted by the constraints on capital decumulation. We might call the range of values to the left of Q the "stability range" of the system.

Models of the type under discussion can produce depressions in two different ways. These possibilities can be illustrated by reference to Figure 9. The economy whose growth properties are represented by the curves shown in Figure 9 is capable of producing sustained growth in demand. If at any point in time the ratio of capital to income is less than the one at point Q in the diagram, income will be rising and the ratio of capital to income and the rates of growth of capital and income will move toward the stable equilibrium values at point P. Suppose now that while the value of $\frac{k}{y}$ lies to the left of Q some adverse shock occurs which temporarily causes income to fall. Suppose, however, that after a time the effect of the shock is over and the behavior equations of the system return to their former positions. Since income has fallen and capital stock has not, the ratio of $\frac{k}{y}$ after the shock will be higher than before. But if after the shock $\frac{k}{y}$ is still to the left of Q, income will begin to rise again after the shock. If, however, the value of $\frac{k}{y}$ after the shock lies to right of Q, income will continue to decline even after the behavior equations snap back to their old position. Eventually income will stop falling because there is a limit to the speed with which capital can be decumulated. Income can begin to rise again after a long period of capital decumulation.

The size of the depression produced by a given shock depends on the initial position of the system relative to point Q, the speed with which income declines under the impact of the shock, and the duration of the shock. These factors are discussed in detail in Chapters 11 and 12. For the moment we need only note the existence of the possibility of producing a variety of types of depressions by applying shocks to a system which is capable of sustained growth.

The other major type of depression occurs when the r_k curve lies above the r_y curve for all positive values of $\frac{k}{y}$, r_y, and r_k. In that case the rate of growth of capital exceeds the rate of growth of income for all values of $\frac{k}{y}$. If the initial value of $\frac{k}{y}$ is sufficiently low, income may rise for a time; but as the ratio of capital to income rises, the rate of growth of income will decline until it becomes negative. Eventually, after sufficient disinvestment takes place, income will begin to rise again, but there will be a long depression in the meantime.

In the remainder of this chapter we shall use the simple model represented in the diagram as a basis for discussing the effect of various types of change in the economic structure on the stable-equilibrium ratio of growth of income. We shall do this by considering the impact of any change in structure on the rate of growth of income and the rate of growth of capital for given values of the capital-output ratio. In general, it can be said that whenever a change in structure raises the rate of growth of income and the rate of growth of capital by the same amount, the equilibrium rate of growth of income will be raised. This can easily be seen by reference to Figure 9. If we shift both the r_y and r_k curves upward by a constant, point P will be shifted vertically upward.

The rate of growth of income will also be raised whenever a change in structure raises the rate of growth of income by more than the rate of growth of capital for values of $\frac{k}{y}$ in the neighborhood of the stable equilibrium.

In cases in which no equilibrium exists, a structural change which raises the rate of growth of income reduces the gap between the r_y and r_k curves. This will either make stable growth possible or at least lengthen the period during which income can grow when the system starts from a position in which the rate of growth of income is positive.

In discussing the effect of structural changes on the rate of growth of income, we shall frequently use the fact that by definition the change in income from one period to the next equals the excess of investment over Robertsonian saving. (The latter term is defined as the difference between consumption in period $t - 1$ and income in period t.)

Our procedure in the remainder of this chapter is as follows: In the next section we consider the effect of various types of lags on the rates of growth of income and capital. Then we consider the effects of changes in the propensity to save, the capital intensity of production methods, changes in rates of profit, credit availability, and governmental expenditure. The discussion to that point is in aggregative terms. We then introduce a multisector model and consider the effect on the growth rate of differences in the rates of growth of demand for different products, population growth, and the introduction of new products.

II. EFFECT OF LAGS

The first step in expanding our model is to introduce lags of more than one period. Instead of rewriting the equations of the model, we shall proceed by discussing the effects of lags on the r_y and r_k curves, i.e., on the relation between the rates of growth of capital and income and the ratio of capital to income.

So far we have assumed that the system operates without any lags of more than one period. In that case the rates of growth of income and capital from one period to the next depend only on the ratio of capital stock to income and not on the history of the behavior of income. When we make consumption depend in part on past consumption, and dividends on past dividends, current expenditure will depend on the levels of capital stock and income over an extended period in the past. In equilibrium the past values of capital stock and income will have a fixed ratio to the values of those variables in the immediate past, and that ratio is determined by the rate of growth of income. The equations for the rates of growth of capital and income will contain r_y and r_k on both sides. The same effect will be produced by the fact that housing demand depends on accumulated savings and therefore on the average income over a number of years in the past. Similarly, investment demand depends not only on the size of the capital stock but also on its age distribution. The age distribution depends on the rate of growth of income. Finally, investment demand depends on expectations about the rate of growth of income, and those expectations depend in part on actual past rates of growth of income.

The effect of those considerations can be illustrated by reference to the consumption lag. Our original equation for consumption out of disposable income was

(1) $$C_t = f(y_{Dt-1}, C_{t-1})$$

Let us assume the linear homogeneous form

(2) $$C_t = Ry_{Dt-1} + qC_{t-1}$$

If we substitute $C_{t-1} = Ry_{Dt-2} + qC_{t-2}$ into equation (2), we obtain

(3) $$C_t = Ry_{Dt-1} + qRy_{Dt-2} + q^2C_{t-2}$$

If we again substitute $C_{t-2} = Ry_{Dt-3} + qC_{t-3}$ in equation (3), we obtain

(4) $$C_t = Ry_{Dt-1} + qRy_{Dt-2} + q^2Ry_{Dt-3} + q^3C_{t-3}$$

Repeating the substitution process n times, we obtain

(5) $$C_t = Ry_{Dt-1} + qRy_{Dt-2} + q^2Ry_{Dt-3} + q^3Ry_{Dt-4} + \cdots + q^{n-1}Ry_{Dt-n} + q^nC_{t-n}$$

Both R and q are less than 1, so that when n is sufficiently large, we may neglect the final term in the series. Suppose that income has been growing at constant rate of r per cent per period. Then

$$(6) \quad y_{Dt-2} = \frac{y_{Dt-1}}{1+r} \qquad y_{Dt-3} = \frac{y_{Dt-1}}{(1+r)^2} \qquad y_{Dt-n} = \frac{y_{Dt-1}}{(1+r)^{n-1}}$$

Equation (5) may be rewritten as

$$C_t = Ry_{Dt-1} + \frac{qRy_{Dt-1}}{1+r} + \frac{q^2Ry_{Dt-1}}{(1+r)^2} + \cdots + \frac{q^{n-1}Ry_{Dt-1}}{(1+r)^{n-1}}$$

It is obvious that for a given value of y_{Dt-1}, C_t grows smaller as the rate of growth r increases.

The rate of growth of income from one period to the next equals the excess of consumption plus investment in period t over income in period $t - 1$. If we reduce consumption for a given value of $\frac{k}{y}$, we reduce the rate of growth of income. The rate of growth of income therefore tends to fall as past rates of growth of income rise. In equilibrium, past and current rates of growth of income are equal.

Suppose that there were no consumption lag, so that C_t is given by $R'y_{Dt-1}$. Given the equations of the rest of the system, we could compute its equilibrium rate of growth. Call it μ_E. Now suppose that consumer behavior changes so that it can be described by the equation $C_t = Ry_{Dt-1} + qC_{t-1}$. Suppose that the static propensity to consume remains unchanged in the sense that at a constant income the proportion of income consumed is R'. In equation (6) put $r = 0$. Then we have from equation (5)

$$C_t = R(1 + q + q^2 + \cdots + q^{n-1})y_{Dt-1}$$

If C_t is to equal $R'y_{Dt-1}$, then

$$R' = R(1 + q + q^2 + \cdots + Q^{n-1})$$

Now if after the change income continued to grow at the rate $\mu_E(\mu_E > 0)$, C_t would be given by

$$C_t = \left[\frac{R(1+q)}{1+\mu_E} + \frac{q^2}{(1+\mu_E)^2} + \cdots + \frac{q^{n-1}}{(1+\mu_E)^{n-1}}\right] y_{Dt-1}$$

which is less than $R'y_{Dt-1}$. If the rate of growth μ_E was achieved when $C_t = R'y_{Dt-1}$, the rate of growth must now fall to a lower level.

The effect of the lag, therefore, is to reduce the equilibrium rate of growth of the system.

Suppose we draw the r_y curve as before, neglecting the lag in consumption. This curve will show correctly the value of $\frac{k}{y}$, at which r_y equals zero. For all smaller values of $\frac{k}{y}$ the true curve which takes account of the lags will lie below the static curve, the distance between them growing larger as $\frac{k}{y}$ grows smaller. The curve which shows the effect of the lag tends to flatten out as $\frac{k}{y}$ grows smaller and r_y grows larger. The relation between the two curves is shown in Figure 10 together with an r_k curve. It is important to note that the reduction in the stable-equilibrium value of r_y is greater than the reduction in the value of r_y at any given value of $\frac{k}{y}$. That is so because the reduction in the level of the r_y curve pushes the stable intersection of the r_k and r_y curves downward and to the right. At the same time the unstable equilibrium is pulled to the left, so that the distance between the two equilibria is reduced considerably.

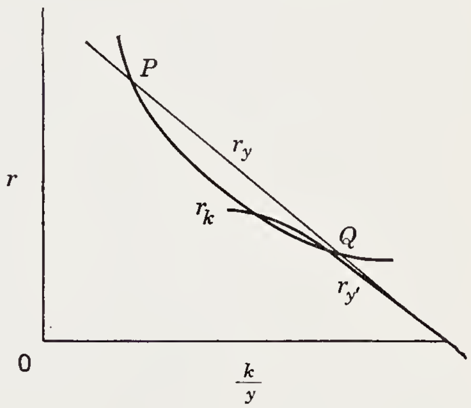

FIGURE 10.

The dividend lag and the lagged response of housing demand to changes in income will have a similar effect on the r_y curve. However, the lag in housing demand will affect the growth of capital stock as well as the rate of growth of income. Shifting the r_k curve downward tends to raise the stable-equilibrium rate of growth and lower the unstable one, thus offsetting in part the effect of the reduction in the r_y curve. The effect of the downward shift in the r_y curve will be only partially offset.

The lag between replacement and depreciation can be regarded as influencing the rate of growth of capital without changing the rate of growth of income directly. It is sometimes assumed that capital disappears at the same rate as it is depreciated in ordinary bookkeeping practice. In some cases capital loses its serviceability gradually, and it becomes profitable to build new capital (other things equal) at a rate equal to the rate at which depreciation charges are accruing. (See the discussion of depreciation of housing above.) In other cases, however, particularly those in which obsolescence is important, bookkeeping depreciation is a poor measure of the effective capital stock. In many cases the probability that a firm will find it profitable to take a piece of equipment out of service when it is, say, from zero to nine years old is very small and then begins to rise rapidly. A firm having a given amount of equipment valued gross of depreciation will therefore invest about the same amount when all its equipment is one year old as when all its equipment is nine years old. If we measure the stock of capital net of depreciation, we get the same rate of investment with two different capital stocks.

For that reason it is appropriate to measure the capital stock gross of depreciation. The annual increment in the stock is then equal to the current amount of gross investment less the amount of gross investment made N years before, where N is the life of the capital. When income is growing steadily the investment in year t is $(1 + r_y)^n$ times investment in year $t - n$. Therefore

$$k_t = I_t(1 + r)\,\frac{1 - \dfrac{1}{(1 + r)^n}}{r}$$

As $\dfrac{k}{y}$ decreases and r_y increases, the curve for r_k rises faster than it would if straight-line depreciation were a proper way of measuring the influence of capital deterioration on investment. Just as in the other cases, the "Domar effect" tends to prevent the rate of growth of income from rising to very high levels.[1]

The interaction between rates of growth and expectations tends to increase the equilibrium rate of growth by comparison with the rate which would rule in the absence of expectational effects. It was shown in Chapter 2 that the amount of investment generated by a given current situation in terms of capital stock and income varies with the ratio of expected future levels of demand to the current level. At the same time it is reasonable to suppose that the ratio of expected future demand to current demand will increase as the observed past rate of growth increases. Just as in the various lag cases considered above, the rate of growth of income enters the equation for investment and therefore enters the equation for the rate of growth of income on the right side of the equality. In the other cases, however, high rates of growth of income have a depressing effect on investment and consumption (by comparison with static equations with the same parameters). In this case the effect is in the opposite direction. The expectational effect tends to reinforce any tendency for the equilibrium rate of growth to rise.

The influence of long-term growth expectations is, however, a limited one. If we suppose that we really have to deal with a stationary situation in which income is rising smoothly and steadily, the expected future rate of growth of income influences investment demand only through its influence on the timing of construction of discontinuous units of plant, and no matter how high the rate of growth becomes, the influence of that consideration is limited by the size of the discontinuities.

If we depart from the assumption of a perfectly steady growth path, expectations become more important. When income is subject to some fluctuation, uncertainty as to future levels of demand becomes important in investment policy. If firms believe that aggregate income is going to grow rapidly on the average, the risk of loss

[1] E. D. Domar, "Depreciation, Replacement and Growth," *Economic Journal*, vol. 63, 1953.

Robert E. Eisner, "Depreciation Allowances, Replacement Requirements, and Growth," *American Economic Review*, vol. 42, 1952.

as a result of future depressions is greatly reduced. But the fluctuations which make expectations about the trend of aggregate demand important also weaken the connection between the observed average rate of growth in the past and the expected future rate of growth. If income rises steadily for a number of years, managements will develop a strong tendency to project that growth rate into the future. But if income fluctuates about on upward trend, the tendency to project the trend into the future will be greatly weakened.

III. SOME COMPARATIVE DYNAMICS

The ultimate purpose of a discussion of the equilibrum rate of growth of income is to gain some insight into the influence of various types of structural change on the rate of growth. It takes considerable time for the system to converge to the equilibrium corresponding to a given set of parameters. Therefore movements in the equilibrium rate of growth are not identical with movements in the actual rate of growth. However, it is generally true that the actual rate of growth will move in the same direction as the equilibrium rate of growth. Indeed, it will quite often happen that the immediate increase or decrease in the rate of growth generated by a parameter change will be greater than the change in the equilibrium rates generated by the parameter change.

An exhaustive analysis of the impact of every kind of structural change on the equilibrium rate of growth of income and on the stability of the system is impossible. A few examples of the effects of the most important kinds of structural change will have to suffice.

A. Propensity to Consume Out of Personal Income

It is obvious that, other things equal, an increase in the personal propensity to consume tends to increase the equilibrium rate of growth of income. It raises the r_y curve without changing the r_k curve, so that the value of r_y at which r_y equals r_k is increased.

For the same static propensity to consume, the rate of growth of income will increase if the lags in the response of consumption to income are reduced. It is therefore possible to achieve the same rate of growth with a relatively low static propensity to consume and short lags as with a high static propensity to consume and longer lags.

B. Dividends and Entrepreneurial Withdrawals

The direct effects of high static dividend pay-out ratios or short lags in the response of dividends to profits are similar to the effects of a high propensity to consume or to short consumption-response lags. However, any change in dividend policy must be accompanied by changes in other aspects of business policy. If high dividend pay-outs are forced by tax measures or stockholder pressure, the resulting increase in consumption might be offset by a reduction in business investment. Indeed, the reduction in investment might exceed the increase in consumption (*a*) because the propensity to consume out of dividends is probably fairly low, (*b*) because firms might borrow less in order to maintain the same relationship between debt finance and equity finance (from retained earnings).

In the long run the restriction of investment would raise profits so as to make the rate of retained earnings per dollar of invested capital about the same as it was originally. Retained earnings would, however, be a lower fraction of income than before because the increase in earnings per dollar of capital takes place as a result of a fall in capital intensity. In those circumstances a forced change in dividend policy would have only a small effect on the rate of growth of income.

If a more generous dividend policy were caused by or accompanied by a reduction in equity yields, finance by retained earnings might be replaced by finance by sale of equities. In that case the increase in dividend pay-out would increase the rate of growth of income.

That would also be true if the increase in dividends were restricted to firms having retained earnings in excess of their own expansion requirements. The increased dividend payments would then cause a reduction in liquid-asset holdings of the firms in question or of the rate at which they acquire other firms.

C. Changes in Rates of Return

Changes in the average rate of return on capital for given $\frac{k}{v}$ may be brought about by increased competition. Marginal efficiency of investment is bound to change at the same time, but when the change in average profits is due to a change in competitive structure, the fall in marginal efficiency will be much smaller than the fall in average profits. Indeed, it is possible for marginal efficiency to rise while average profits fall. If, for example, entry into certain industries becomes easier, investment at a given value of $\frac{k}{y}$ will rise and profits will fall $\left(\text{at a given value of } \frac{k}{y}\right)$ because of greater short-run competition. An increase in investment in relation to $\frac{k}{y}$ and a fall in retained earnings are sure to increase the rate of growth of income.

D. Availability of Credit

Changes in the availability of credit influence both the rate of growth of income and the rate of growth of capital stock. A reduction in mortgage down payments tends to increase the rate of growth of income (at given capital-income ratio) by reducing the lag between change in income and changes in the demand for single houses. At the same time it increases the rate of growth of the stock of houses. Change in the availability of credit to other types of enterprises, e.g., less restrictive credit rationing for customer loans or the use of term loans instead of short-term loans, will tend to increase the rate of investment corresponding to a given level of capital stock and income. Such changes may also tend to reduce the rate of business saving. Easier credit conditions therefore raise the r_y curve and at the same time shift the r_k curve upward.

In general, one would expect that an upward shift in both curves would tend to raise the equilibrium rate of growth of capital while leaving the equilibrium capital-income ratio more or less unchanged.

E. Capital Intensity

The capital intensity of production methods is obviously important in determining the rate of growth of income. Strictly speaking, the term "capital intensity" has no meaning except in the case where substitution is impossible, because the amount of capital actually employed to produce a unit of output depends on the relative prices of capital goods and other inputs and on the cost of raising capital, as well as on the technical production function. However, it is clear that in some industries all methods of production are more capital-intensive than those available in some other industries. Similarly, there are some periods in which the relevant production possibilities are generally more capital-intensive than in others.

Those who use acceleration-principle models tend to assume that an economy whose methods of production are highly capital-intensive will grow at a higher rate than an economy whose methods are less capital-intensive. In general, that is probably true, but the argument is not as simple as it appears. The rate of growth of income depends on the difference between the rate of investment and the rate of planned saving. A considerable part of saving is done by business firms, and saving varies with profits. To determine the effect of differences in capital intensity on the rate of growth of income it is necessary to consider the effect of the differences on profits and business saving as well as their effect or investment. A more capital-intensive technique leads to a higher rate of growth of income than a less capital-intensive technique provided that the more capitalistic technique is not associated with a higher rate of return on capital than the less capitalistic technique. When the two techniques produce the same rate of return on capital, the ratio of borrowing to profits will be the same in both techniques at any given rate of growth of income, but the ratio of borrowing to income will be higher with the more capital-intensive technique. That will produce a higher rate of growth of income.

Except to the extent that high capital intensity fosters monopoly, there is no systematic reason why high capital intensity should be associated with high profits, so that with that qualification we can suppose that changes in technique leading to higher capital intensities do tend to increase the rate of growth of income. But that need not be true in every individual case.

It should be noted that the above argument is based on considerations relating to differences between capital intensities. In the transition from a low capital intensity to a higher one the rate of growth of income will almost always be increased.

F. Effect of Technical Progress on Demand

The systematic effect of technical progress on the growth of demand is not likely to be very great (if we exclude from the category the effects of technical progress on the "ceiling" problem). Let us suppose that technical progress takes the simple form of reducing labor-input requirements per unit of output with given capital inputs per unit output. In terms of an isoquant diagram, that amounts to saying that, if the rate of technical progress is 2 per cent per year, the scale of the labor-input axis expands by 2 per cent per year. One unit of labor is represented by a progressively longer distance along the axis. Meanwhile the capital axis and the isoquants themselves remain unchanged. Suppose that the change in labor efficiency generated by better equipment design extends equally to all industries, including the capital-goods industries, and that wages rise at the same rate as productivity. In this situation, capital goods and other prices will remain constant. Since wages are rising, capital-goods prices fall relative to wages. But capital-goods prices retain a constant ratio to "efficiency" wages. One dollar's worth of capital exchanges for a progressively smaller amount of labor but for a constant amount of labor efficiency. As a result the amount of capital used per unit of output (at a given imputed interest rate) remains unchanged, while the amount of labor per unit of output falls and the amount of capital per worker rises. All the behavior functions remain unchanged. This type of neutral technical change has no effect on the rate of growth of either real or money income. There will be no technological unemployment if the rate of growth of income is as large as the sum of the rate of growth of labor supply and the rate of growth of productivity. The adjustment between demand for and supply of labor is discussed in Chapter 8.

It is not necessary, of course, that technical change be neutral. Technical change may cause changes in the relative prices of capital goods and labor, or it may cause changes in the position and shape of the isoquants. There is, however, no obvious reason to assume any particular type of change in relative prices, isoquant shapes, or positions. As we have already seen, there is some reason to think that an increase in capital intensity (a shift upward to the left of the point at which an isoquant for given output reaches a given slope when capital is measured vertically and labor horizontally) tends to raise the rate of growth of the system. The effects of changes in the relative prices of capital goods or in the shapes of isoquants can be analyzed only by using explicit production functions in the model.[2]

G. Government Expenditure and Taxation

It is, of course, impossible to give an adequate treatment of the effects of government expenditure and taxation in a general work like this. But government activity is so important these days that it is necessary at least to indicate the points at which it impinges on the income-generation process.

[2] An alternative formulation of unbiased technical change is the following. Suppose that each industry has a production function of the Cobb-Douglas type. $O = kL^aK^b$, where O is output per period, L is labor input per period and K is capital stock. Suppose that technical improvement results in a progressive rise in k in each industry, leaving a and b constant. The isoquants of the production functions then retain the same shape, but the output number attached to each isoquant rises as time passes. Now suppose that money wages rise with k. If the rate of profit on money capital invested remains constant, the prices of all goods will fall. (For if the prices remained unchanged, capital invested per unit of output would fall in proportion to the rise in k, while the margin between wage costs per unit and price remained unchanged.) If wages rise and capital-goods prices fall, there will be substitution of capital for labor. More capital per worker will be used, but whether more or less capital per unit of output will be used depends on the elasticity of substitution between capital and labor.

If the elasticity is large enough, the ratio of capital to output will remain constant, and the effect will be about the same as in the other case.

The effects of the government-expenditure side of the picture are simple enough if we regard government expenditure as arbitrarily given. In our equation for income growth we can simply add government expenditure to private-consumption expenditure and private-investment expenditure. Then the rate of growth of income will be increased by the amount of government expenditure expressed as a percentage of income. The increase in the percentage rate of growth of income will be the same regardless of the capital stock and income position at the time when the expenditure is injected.

The equilibrium rate of growth of income will also be changed by government expenditure. In Figure 11 the curve marked r_k represents the rate of growth of capital stock as a function of the capital-income ratio. The curve marked r_y shows the rate of growth of income when there is no government expenditure. The point P is the equilibrium point in terms of rate of growth of income and capital-income ratio. If we now introduce government expenditure which is some given percentage of initial income, the r_y curve shifts upward by the amount of government expenditure as a percentage of income to r_{y2}. The new equilibrium rate of growth is P_2. But if the government expenditure is a constant amount, it will be a smaller and smaller percentage of income. r_{y2} will therefore shift slowly downward, and after infinite time will coincide with r_{y1}.

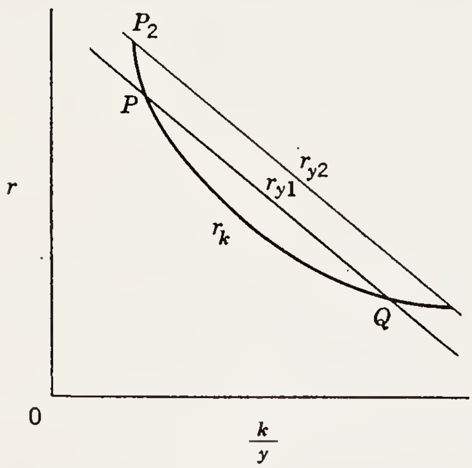

FIGURE 11.

It will be noted that uncompensated government expenditure has a very large, if not infinite, "multiplier." That is so because it influences investment as well as consumption expenditure.

The analysis of the effects of taxation is much more complicated even if we keep it as simple as possible. In the simplest case, taxation consists of lump-sum poll taxes. Suppose everyone has a propensity to consume of a. Then a government expenditure of G financed by poll taxes will raise the rate of growth of income by $\frac{(1 - a)G}{y}$ percentage points, where y is aggregate income. The rate of growth will then increase as the system moves toward its new equilibrium, then decrease as the government expenditure becomes a smaller fraction of income.

Other types of taxation will have similar effects, but the exact magnitude of the effect will depend on the propensity to save of those subject to taxation. However, when taxation restricts personal saving, it may have some side effects on investment, depending on the money-market situation.

Taxation of business income has much more complicated effects. Because investment is influenced by the amount of retained earnings available to business firms, corporate income taxation will reduce investment expenditures. Corporate income taxation will also reduce dividends, so that consumption will fall by the dividend–pay-out ratio times the amount of the tax times the propensity to consume of dividend recipients. The reduction in investment that results from the tax is not certain, but it is probably a large fraction of the reduction in retained earnings. The direct result of an increase in government expenditure financed by corporate income taxation will be an increase in the rate of growth of income. The increase will equal the expenditure as a fraction of income times the dividend–pay-out ratio times the propensity to save of dividend receivers minus the reduction in investment resulting from the expenditure. But because investment as a function of income and capital is now smaller than before, the rate of growth of capital function is shifted downward. The equilibrium rate of growth of income will shift up farther than in the other cases. The equilibrium capital-income ratio will be lower than before. As a result corporate profits before tax will rise as a fraction of income, and at least part of the corporate income tax will have been shifted. Of course, the practical results of corporate income taxation will depend in large measure on the exact terms of the tax law and the accompanying regulations.[3]

[3] The argument given above assumes that the prices were such as aximizetmo

So far we have treated government expenditure and taxation as exogenously given. For Federal expenditures, that is not inappropriate. One might take all state and local expenditure and taxation as a function of income, but it is particularly important to do so in the case of schools, highways, and other capital expenditures. In regard to those expenditures, state and local governments are somewhat in the position of business firms. Their borrowing power is limited by their ability to raise revenues. Their ability to raise revenues depends on the incomes of the taxpayers and their willingness to pay taxes for interest and bond amortization. The willingness of the taxpayers to be taxed for those purposes depends on the existing stock of public-capital assets.

It would not be foolish, therefore, to regard state and local governments as part of the business sector selling the services of public capital to taxpayers and investing when the ratio of public capital to income is sufficiently low. That approach has some significance in an income-generation scheme because state and local governments finance a very large part of their capital expenditure by borrowing. A shift of public taste toward public capital can have a substantial effect on the rate of growth of income.

IV. MULTISECTOR EQUILIBRIUM

An aggregate model can be used to gain an understanding of the nature of the interaction of capital growth and income growth, but it is essential to take some account of differences among the sectors of the economy if a realistic explanation of observed movements of income is to be given.

Even if final demand for each class of product varies in proportion to income, the relations among aggregate capital stock, income, investment, and consumption will not be unique unless the behavior parameters for all the sectors are exactly the same. When the be-

---

short-run profits before the imposition of the tax. If profits have been held below the short maximum for fear of entry or some similar reason, all or part of the tax will be shifted at once. That may also be true if prices initially are set at a somewhat arbitrary level because of the difficulty of determining the optimum price.

havior parameters of the sectors differ from one another, the rate of investment associated with a given aggregate capital stock will vary with the composition of the stock, i.e., with its distribution among industries. There is no way to average the parameters of the different sectors. The only proper procedure, then, is to analyze a multisector system. That creates a great deal of difficulty from an empirical point of view, but from a theoretical point of view the situation is not so bad.

A multisector system in which final demand is proportional to aggregate income has equilibrium properties analogous to those of a single-sector system. Let us consider the multisector system analogous to the simple unlagged system with which this chapter was introduced. There we had

$$\begin{aligned} I_t &= \alpha y_{t-1} + \beta k_{t-1} \\ C_t &= a y_{t-1} + b k_{t-1} \end{aligned}$$

In a multisector system there is a different set of parameters for each industry. Denoting the parameters applying to the ith industry by A_i, β_i, etc., let l_i be the proportionality factor between aggregate income and an index of the level of demand for the ith industry. We can then write

(1) $\quad I_t = \alpha_i l_i y_{t-1} + \beta_i k_{it-1}$

where k_i is the capital stock of the ith industry.

(2) $\quad C_t = a y_{t-1} + \Sigma b_i k_{it-1}$

(3) $\quad y_t = \Sigma I_{it} + C_t = (a + \Sigma \alpha_i l_i) y_{t-1} + \Sigma(\beta_i + b_i) k_{it-1}$

(4) $\quad r_y = \dfrac{y_t - y_{t-1}}{y_{t-1}} = (a - 1 + \sum \alpha_i l_i) + \sum (\beta_i + b_i) \dfrac{k_{it-1}}{y_{t-1}}$

For the ith industry the increase in capital stock is given by $\alpha_i l_i y_{t-1} + \beta_i k_{it-1}$, and

$$r_{ki} = \frac{k_{it} - k_{it-1}}{k_{it-1}} = \alpha_i l_i \frac{y_{t-1}}{k_{t-1}} + \beta_i$$

The equilibrium condition for the multisector system is that the rate of growth of capital for any sector should equal the rate of growth of aggregate demand.

There will be many solutions to the resulting system of equations. Suppose there exist real and positive values of r_y which satisfy the equations. Now consider the largest of these real roots. It will be associated with relatively small equilibrium values of the $\frac{k_i}{y}$'s. It will almost surely be stable from above. Suppose we select initial values of the $\frac{k_i}{y}$'s such that income initially grows faster than the equilibrium rate (corresponding to the largest real value of r_y). *Ex hypothesi*, most types of capital will grow at a higher rate than income, and that will tend to make income slow down.

The capital income ratios for some sectors will be nearer the equilibrium value than others, and the rate of approach to equilibrium will be greater in some sectors than in others. It is therefore possible that some sectors will overshoot the mark. But the capital-income ratio cannot increase very rapidly in any sector. It is not very likely, therefore, that any sector will acquire a capital-income ratio which exceeds the equilibrium ratio to a serious extent. It seems likely, therefore, that the rate of growth of income will converge to its equilibrium value with no more than minor oscillations.

If the capital-income ratios are all slightly in excess of their equilibrium values, convergence to the equilibrium rate of growth is still possible, for then all types of capital will accumulate at rates which are below the rate of growth of income. Capital-income ratios will fall, and the rate of growth of income will rise.

We can see by analogy from the single-sector model that when the average of the capital-income ratios becomes sufficiently large, capital stocks will accumulate more rapidly than income, so that the system moves away from its equilibrium position. The rate of growth of income depends not only on the aggregate size of the capital stock but also on the correlation of the amounts in the different sectors with the sector investment and saving functions. Capital accumulates at different rates in the different sectors. Therefore the correlation will change as the different sectors accumulate or decumulate capital.

If the ratio of one type of capital stock is above the equilibrium and if investment in that sector is very sensitive to the capital-income ratio, there may be enough disinvestment in that one sector to cause a fall in aggregate income. But capital in the sector in question will fall faster than income falls, so that the capital-income ratio in the sensitive sector declines. If capital stocks in other sectors are not excessive, the rate of growth of income can move back toward its equilibrium value after the adjustment.

However, if capital stocks are generally above their equilibrium value, income will decline and a long depression will ensue before the system can reach its equilibrium growth rate.

The set of all possible combinations of sector capital-income ratios can be divided into three subsets. If the initial situation corresponds to an element of one subset, the system will converge to equilibrium with minor oscillations in the rate of growth. For initial conditions corresponding to the second set of combinations of ratios, the system will converge to equilibrium, but income will actually fall at least once before it begins to grow steadily. But the depression will be a short one.

For initial conditions corresponding to the third set, income will fall and there will be a general reduction in capital stocks before equilibrium can be reached.

The reader will see that this is a generalization of the stability-range concept used in connection with the single-sector model.

V. GROWING INDUSTRIES AND AUTONOMOUS INVESTMENT

We may learn something about the real world by studying stationary systems, but the real world is certainly not stationary. To understand the real world we must take account of changes in production methods, the development of new products, growth of population, and the opening up of new areas of settlement. Those phenomena influence the economic system, but their occurrence is not primarily dependent on economic conditions. They are usually classified as exogenous events operating on the economic system. In some treatments, technical change, new product development, and the opening of new territories are said to produce "autonomous investment." Total investment is then obtained by adding induced investment to exogenous investment. At the same time it is often assumed implicitly that the occurrence of autonomous investment has no impact on the rate of gross saving. Neither of these positions is tenable.

Each of the four classes of events mentioned above has a direct positive effect on investment in some sectors. Each of them may also have a negative effect on investment in other sectors or a positive effect on saving. In large measure these two sets of effects offset each other.

The development of new products obviously generates investment in plant and equipment to produce the products in question. To some extent at least, the sales of the new products are at the expense of sales of old products. The positive effect of investment in plant for new products is at least partially offset by reductions in the rate of investment in facilities to build old products. The offset need not be complete, however. The new products may be more capital-intensive than the old ones, or the new industry may borrow a greater proportion of its capital than the old industries would have borrowed. In addition, the time distribution of investment may be changed by the appearance of new products. Essentially the same argument applies to the development of new sources of supply, whether they are on the frontier or elsewhere.

Improvements in production methods influence investment directly by speeding up replacement. At the same time they influence gross saving by raising depreciation charges. There is obviously no one-to-one correspondence between any particular technical improvement and any depreciation charge. But it seems equally obvious that depreciation charges will be higher in a world of rapid obsolescence than they would be if equipment were replaced only when unserviceable or subject to high maintenance cost. The direct influence of technical change on investment is largely offset, therefore, by its influence on gross saving. Once again, however, change does have some net influence on the rate of growth of demand. It influences the size of the lag between depreciation and replacement, and it influences the time distribution of investment.

Improvements in production methods, new products which are inputs for other products, or new lower-cost sources of raw-material supplies have another and quite different kind of influence on investment. They all increase the output potential of the economy (i.e., they shift the transformation curve outward).

In most cases they reduce the chance that capital accumulation will drive down the returns on investment by increasing the relative scarcity of other resources. That consideration is important, but it has only an indirect effect on investment. In neoclassical theory, the development of techniques which reduce capital inputs less than labor (or other noncapital) inputs at given prices tends to cause price adjustments which increase the returns from investments. But when the velocity of money can vary so that total expenditures can vary, it is not at all clear that unemployment leads to falling real wages and rising returns on investment. It is obvious that labor shortages may exercise an adverse influence on investment, either directly or through their influence on prices and the availability of credit. But once the labor supply becomes slightly larger than the demand, at going rates, further increases in labor supply (or reduction in demand) do little to encourage investment. It seems useful to ignore, for the moment, the impact of technical change on factor scarcity. That problem can be dealt with more appropriately when we deal with ceiling theories of the trade cycle in the next chapter.

Population growth comes into the picture in several ways. Like technological change, it reduces the danger that the growth of income will be checked by labor shortages. We shall ignore that effect for the moment. Population growth has a systematic influence on demand in two ways. For a given rate of increase of aggregate income the rate of increase of per capita income falls as the rate of population growth rises. As a result the demand for commodities with income elasticities of demand below 1 grows more rapidly as the rate of growth of population increases. Housing has an income elasticity of less than 1. Since housing is the most capital-intensive of all commodities, a high rate of population growth (for given growth of total income) tends to raise total investment. At the same time a high rate of growth of population tends to reduce the average lag between consumption and expenditure by reducing the rate of growth of per capita income. In those two respects, then, an increase in the rate of growth of population tends to raise the equilibrium rate of growth of aggregate demand. It is not clear, however, that a high rate of population growth tends to increase the rate of growth of per capita income.

With those general considerations in mind we can now examine the impact of technical change and new-product development in more detail.

A. Growing Industries

In our equilibrium model we assumed that the demand for every product shifts upward at the same rate as the GNP. That was a useful simplification, but everyone knows that the demand for some products grows much more rapidly than the GNP, while the demand for others grows much more slowly. Those differences arise from a number of causes: (1) Other things equal, demand for products with low income elasticities of demand will grow more slowly than GNP, while demand for products with high income elasticities of demand will grow more rapidly. (2) Differences in rates of technical progress will change relative prices and cause differential rates of growth of output. (3) Producers or consumers may substitute relatively new products or materials for old ones, so that demand for the new products grows relative to the GNP, while the demand for older products declines relative to GNP. Some emphasis should be placed on the term "relatively new" because sales of some products (at given price and income) may continue to rise faster than income for many years after the product first appears.[4]

[4] Of course, the relative magnitude of the sales of different industries will not be invariant to changes in the rate of growth even if there are no changes in the structure of consumer expenditure or in the interindustry input-output coefficients. The ratio of investment to aggregate income will vary with the rate of growth of aggregate income. However, that consideration is not likely to be of very great importance. In the first place, gross investment will not vary as much as net investment with changes in the rate of growth. Moreover, only a few industries derive all their demand from investment expenditures. The effect of changes in the rate of

What is the influence of these differential rates of growth? It is shown in the next section that differences in rates of growth of demand for individual products influence the rate of growth of aggregate demand when either of two conditions is satisfied. The conditions are: (1) There is a correlation between the rates of growth of demand for individual products and the behavior parameters of the industries producing them, e.g., capital intensity of production methods. (2) The redistribution of demand is anticipated and is accompanied by significant indivisibilities in production methods.

For convenience the two cases can be considered separately. In the next section we consider the effect of differential rates of growth of demand, neglecting the effect of anticipations and indivisibilities. We can then consider the additional effect of those factors.

1. *Differences in Growth Rates: No Anticipations*

The consequences of changes in the distribution of demand among the sectors can be dealt with in the following way. Over any short period we can take the shares of demand going to the different sectors (represented by l_i in the last section) as given. We can apply the same equations as developed in the last section but change the equilibrium condition. If each sector is to maintain constant capital intensity, while aggregate demand grows, it is necessary that the growth rate of each sector's capital stock should equal the rate of growth of aggregate demand plus the rate of change in its share of aggregate demand (positive or negative). Let the rate of change of share of aggregate demand for the ith sector be P_i. Then the equilibrium condition for the system is $r_{ki} + P_i = r_y$ for every i.

---

growth on the relative sales of different industries will be still smaller for that reason. Gross investment is, in any case, a relatively small part of total income, so that the variations in question are not very large in any case. Finally the industries which derive a large part of their demand from investment are not systematically very different from the others. Some are capital-intensive, others less so. In a model as crude as the one used here, there is probably little to gain by attempting to take account of changes in the relative sales of different industries which result from changes in the rate of growth.

The effect of deviations of the P's from zero can be indicated in the following way. Suppose that there are only two industries and that their initial outputs are equal (in value terms). Now suppose that P_1 becomes positive, while P_2 assumes a negative value of the same absolute magnitude. From the equilibrium equations

$$r_{k_1} = r_y + P_1 \qquad \text{and} \qquad r_{k_2} = r_y - P_1$$

The rate of growth of capital stock for industry A will be greater than the equilibrium rate of growth of income, and the rate of growth of capital stock for industry B will be less than the equilibrium rate of growth of income. If the two industries are exactly the same except for the difference in rates of growth, the resulting effect on the equilibrium rate of growth of income will be small. Since $r_{k_1} = \alpha_1 \dfrac{1}{\frac{k_1}{y}} - \beta$ has to increase, $\dfrac{k_1}{y}$ has to decrease, and for the same reason $\dfrac{k_2}{y}$ has to increase. But the effect of those changes will approximately cancel out.

If industry A is more capital-intensive than industry B and borrows the same proportion of its capital requirements, the increase in the relative growth rate of industry A will tend to increase the equilibrium rate of growth of income. For suppose that income continued to grow at the old rate r_y^0. In equilibrium

$$r_{k_1} = r_y^0 + P_1 \qquad r_{k_2} = r_y^0 - P_2$$

But $\dfrac{k_1}{y} > \dfrac{k_2}{y}$. Therefore the rate of growth of total capital stock must be greater than it was before the redistribution of demand, and therefore greater than r_y^0. If the same proportion of the increase in capital stock is financed by borrowing, the ratio of borrowing to income must rise, and that will increase the rate of growth of income.

Any shift in the distribution of rates of growth toward the more capital-intensive industries will increase the aggregate rate of growth of income, provided that the industries whose shares increase borrow at least as high a proportion of capital requirements as those whose shares decrease. The relatively high rates of growth of transportation and public-utility industries contributed to growth in the nineteenth century, therefore, not only because they are capital-intensive but also because they borrow a particularly high proportion of their capital requirements.

2. *Effect of Anticipations and Indivisibilities*

The argument given in the last section covers a wide range of cases of differential rates of growth. There are many cases in which indivisibilities are not very important. There are also many cases in which the rate of growth of demand is only a little higher than the rate of growth of GNP. In those cases, anticipations will have only a very small effect on investment. Moreover, the expectational effects may cancel out. Investment in industries growing faster than GNP may be stimulated by confidence in continued growth, but the reverse will be true of industries growing more slowly than GNP.

There are, however, a number of cases in which anticipations and indivisibilities will have an important effect on the rate of growth of aggregate income. Take the case of a wholly new product, e.g., a new chemical or plastic. The firm or firms which have developed the product or acquired the rights can estimate the cost of production at various rates of output. At the same time, market analysis will indicate the prices at which the new product can compete effectively with various old products for which it is a substitute. Suppose that there are substantial economies of scale up to a certain rate of output; i.e., the long-run cost curve slopes downward fairly steeply, then levels off. Suppose that at the same time there is a considerable elasticity of demand. At high prices the product can be sold only for a few specialized uses in which certain special qualities are important. At lower prices it will become competitive with other products used in much larger volume. The firm may decide to introduce the product at a high price and produce on a small scale. But if it wishes to take advantage of its technical lead, it will have to build a large plant and establish itself in the high-volume market before other firms do so. That may involve building a plant costing several million dollars, even though it may take several years before the plant can be worked at capacity. The success of the enterprise depends on taking business away from other firms, but the investment expenditure must be made a long time before other firms lose any business. In that case the firm making the investment expects to gain business, but the other firms do not expect to lose any. Meanwhile the investment expenditure of the innovating firm may influence the movement of income in such a way that other firms in the aggregate will not in the end lose any business. Particular firms will, of course, lose business, but others may gain as much or more.

To illustrate the point, let us consider the impact of a short burst of innovation on an otherwise stationary economy. Let us suppose that we have an economy in which each sector's demand grows at the same rate as aggregate demand. Suppose this system has an equilibrium rate of growth of 2 per cent per year and an equilibrium capital income ratio of 2.5. Now imagine that a large number of innovations occur together. In the aggregate the firms introducing these innovations spend 2.5 per cent of GNP on building new plant. They anticipate that after several years the value added (or contribution to GNP) of the product produced by the new plants will aggregate 1 per cent of GNP at the time of building. That implies an average capital coefficient of 2.5 (on a value-added basis).

Suppose that GNP was 100 in the year before the new project was launched. Without the projects it would have risen at the end of one year to 102 and at the end of five years to slightly over 110. Because of the extra investment generated by the new projects, income rises to 104.5 at the end of the first year. Most of the initial increase in sales will go to firms producing capital goods. Because of the fluctuating nature of their business those firms will not be induced to invest on the basis of a short-run increase in demand. However, to the extent that they retain profits from the extra sales their liquid-asset position will improve. Their investment will eventually be influenced by the temporary increase in sales. For that reason, and because of the lags in consumption and dividend and investment responses throughout the system, the respending of the original-investment outlay will proceed slowly. The non-innovating firms will tend to gain business gradually as the effects of the original investment work themselves out. Meanwhile they will also tend to lose sales to the innovating firms.

The net effect of the innovation can be broken down into two parts. First, suppose that the new plants generate no sales for *n* years and then work at capacity. Then the investment associated with the innovations raises the sales of the non-innovators. If the system was in equilibrium before, income and capital stock were rising at the same rate. Now income has been increased and capital in the non-innovating industries has not increased.

The situation is the same as though there had been a one-time uncompensated increase in government expenditures. The ratio of income to capital is raised above its equilibrium value so that income tends to grow at a more than 2 per cent rate for a time. In addition, the base on which the rate of growth is calculated is also higher. The absolute rate of growth will therefore be increased by the innovation even when the percentage of growth falls back to 2 per cent. After a number of years, then, income will be substantially above the level which would have been reached without the innovation. When the new products begin to sell, the effect is the same as though the capital stock had been increased without any investment. Income will now begin to grow at a rate below 2 per cent, but from a base increased by the income-generating effect of the earlier investment. When the system finally returns to its 2 per cent equilibrium, income will be higher than it would have been without the innovations. The amount of gain in income depends on the length of time between the investment generated by the innovations and the sales from the resulting new capacity. If there is no lag, investment in the new industries simply replaces investment elsewhere. If the lag is very long, the innovations will increase income by as much as the uncompensated increase in expenditure.

Of course it is necessary to add or subtract the effect of difference in capital intensity between the new industries and the old ones.

One additional factor has to be considered. It has been assumed that investment to produce new products is a net addition to total expenditure in the period in which it takes place. But many new products are launched by existing firms. If the firms raise outside capital for the purpose, that makes no difference. If not, it must be supposed that the projects compete with other investment, with dividends, or with additions to financial assets. If investment in new products competes successfully with additions to financial assets, a new addition to total expenditures results. But if it reduces dividends, the net addition to expenditure is smaller because only part of the dividends are saved. Finally, firms may restrict their investment in other lines in order to launch new products. Then one type of investment substitutes for another.

It seems clear that a large part of the direct demand-generating effect of autonomous investment is offset indirectly. Nonetheless, autonomous investment may make a substantial contribution to the rate of growth of demand. A fairly substantial proportion of investment today comes under the heading of autonomous investment, and in the nineteenth century an even greater proportion probably fell into that category.

So far we have talked entirely in terms of completely new products, but a much wider range of investments has analogous effects.

1. When production of an item is strongly market-oriented, e.g., public utilities, each local plant is built with the expectation of future growth of demand. Its construction has the same effect as though the plant were built for a really new product regardless of the length of time since the first such plant was built. It took many years to build the railway and public-utility networks, but each new enterprise in a new locale had the same effect as though a completely new product had been introduced. The same reasoning applies to cases in which market-oriented service facilities, e.g., stores, agricultural processing plants, warehouses, are built ahead of demand in an area in which demand is expected to expand.

2. New plant has often been built because it appears to be possible to earn a profit at the going market price even though demand for the product has not increased. Such possibilities appear as the result of changes in technique, the opening of new areas to agricultural development, or simply because of recognition of latent possi-

bilities. We must include in this category the development of farms in the Midwest during the last half of the nineteenth century as well as the more recent movement of the textile industry to the South.

3. For a variety of reasons (see the discussion of integration and diversification in Chapter 4) firms may undertake the production of goods already being supplied by other firms. It may take them some time to establish themselves in the markets they are trying to penetrate. In that case their investment has much the same effect as though they had introduced new products.

In general, it may be said that situations in which firms are induced to build ahead of demand have the effect of raising the ratio of investment to income for a given ratio of capital to income. If developments leading to building ahead of demand arise continuously, so that there is a constant amount of excess capacity of this type, the effect is the same as that of a permanent upward shift in the investment function. Such a shift will, of course, raise the rate of growth of income.

It is not possible to make any numerical estimate of either the net or the gross amount of investment falling into the categories just discussed. It does seem likely, though, that in recent years a substantial proportion of industrial investment has been generated by the factors mentioned above. In the years between the Civil War and World War I the proportion was probably even larger.

VI. SUMMARY AND CONCLUSIONS

Our objective in this chapter was to try to gain some perspective on the nature of the processes which cause aggregate demand to grow. At the same time we have tried to show how different types of structural change tend to speed up or retard the growth of demand. To do that we began with a very simple view of the economic system and then introduced more and more complications.

We began by neglecting all the long-term dynamic factors in the system—population growth, technical changes, new products, regional development, and so on. It was argued that, even without taking account of those factors, income is determined by a dynamic process. The essential aspect of this process is the systematic interaction between capital stock, income, investment, and saving. The capital stock and income of one period determine the rates of saving and investment for the next. The rate of investment determines the change in the capital stock. The difference between the rate of saving and the rate of investment determines the change in the income. The new capital stock and income determine the investment and saving for the next period, and so on.

It was shown by means of a very simple model that such a process can produce a steady growth in income. But to do that, the behavior equations must have parameters whose values lie in a certain range, and the initial values of the variables must be in a certain range.

If the parameters of the system lie outside certain boundaries, sustained growth will be impossible. Given appropriate initial conditions, income may grow at first, but the rate of capital accumulation will exceed the rate of growth of income. The resulting increase in the ratio of capital to income will reduce investment, and the rate of growth of income will decline until it becomes negative. There may be an eventual recovery, but not until after a long period of capital decumulation.

But even if the parameters of the system are such as to make growth possible, income will not actually grow unless the initial values of the variables are appropriate. If initially the ratio of capital to income exceeds a certain figure, capital will grow faster than income and a depression will ensue. We may therefore have a depression either because the system is incapable of growth or because it starts from the wrong initial condition. In this chapter we have concentrated attention on the analysis of factors which determine whether growth is possible or not. The initial-conditions problem is discussed in the next two chapters.

The results of a simple aggregate model with a single lag are of no value in themselves. They are interesting only because analogous results hold for more complex systems.

Having quickly exhausted the possibilities of the simplest model, we introduced lag elements into the picture. It was shown that most of the lags in the system tend to slow down the growth rate of the system. Distributing the response of consumption or housing construction to changes in income reduces the rate of growth of income and may make sustained growth impossible. Even if sustained growth is possible, the existence of lags makes a cyclical movement about an upward trend possible. Consideration of that possibility is postponed to the next chapter, where it is shown to be relatively unimportant.

After some discussion of the effects of various types of structural change on the rate of growth of income we considered a multisector model in which demand for each sector is proportional to GNP. It was shown that propositions analogous to those developed for the simple aggregate model hold in the disaggregated case.

A more interesting result of disaggregation emerges when some industries grow faster than others. It then appears that if the industries which grow more rapidly than GNP are relatively capital-intensive or borrow heavily, the equilibrium rate of growth will be increased. That consideration is important because the relative rate of growth of demand for housing depends on the rate of growth of population. It also lends a special importance to the growth of public utilities, which are both capital-intensive and dependent on borrowing.

Finally, it was shown that the development of new products may cause firms to build ahead of demand. The rate of growth of income will increase with the rate of development of products which cause firms to build capacity ahead of demand.

Without specifying the numerical values of the parameters in the equations one cannot say anything very precise about the conditions required to make sustained growth possible. We know that, given the right combination of behavior relations, rates of growth of population, rates of technical change, and so on, sustained growth of income is at least theoretically possible. We know perhaps a little more than that. Prior to 1914, income had grown for a long time with occasional interruptions.

It is argued in the next chapter that the interruptions to the growth process were caused by shocks such as speculative booms or monetary crises. That implies that if there had been no monetary crises and if the development of innovations had proceeded smoothly (instead of following the erratic course of speculative activity), income could have grown steadily. In the period between the end of the Civil War and 1900 there would have been less investment during the booms and more during the depressions if there had been no speculative activity. The change in timing might have reduced the average rate of growth somewhat, since there would have been less building ahead of demand. But it is hard to believe that such a change in timing would have made the system incapable of growing steadily.

If that is so, then it can be said that sustained growth is not only theoretically but also practically possible. (But that statement is true only if the system is sufficiently stable to adjust itself to limitations on growth imposed by resource ceilings. It is shown in the next chapter that that is probably the case.)

However, while sustained growth is a practical possibility, it is by no means a certainty. There does not seem to be anything about the nature of a capitalist economic system that tends to produce a set of behavior relations consistent with sustained growth. It is true that there are certain links among the parameters of the system which widen the range of possible situations in which growth can occur. For given values of the other parameters the growth rate tends to fall with the rate of population growth. But there would probably be somewhat less saving with a low rate of population growth than with a high one (after a rather long period of adjustment). That is so because the rate of net saving for retirement is related to the rate of growth of population and because the purchase of houses induces saving. Similarly, if there were less innovation, corporations might invest less and also save less.

But those connections among the parameters are probably not strong enough to guarantee that every capitalist system should be capable of producing sustained growth. It may be said that it is a remarkable coincidence that out of all conceivable savings functions, profits functions, and so on, the ones which came into existence in the nineteenth century were so well adapted to the rates of population growth, development of new industry, and other characteristics of that century. One can only say that, if they had not been, capitalism would never have survived so long. There is no built-in adaptive mechanism which makes sustained growth always possible.

It does appear, in fact, that, in spite of some links between population growth, innovations, and the propensity to save, the possibility of sustained growth is very much dependent on population growth and on the relatively rapid growth of particular industries.

Income cannot grow unless investment exceeds Robertsonian savings in each successive period of time. That cannot happen unless someone is prepared to borrow more than the Robertsonian savings of each period. Most businesses other than public utilities are reluctant to borrow to a very large extent. It is sometimes implied that the tendency of most businesses to rely on internal finance is a recent one, but it is probably true that most nineteenth-century businesses depended more heavily on internal finance than do the corporations of today.

If population were stationary and demand for electricity grew only as fast as gross national product, it seems very unlikely that income could grow at all.

Some other industry may some day play the role of electric utilities, but it is true that capital-intensive externally financed utilities (railroads, street railways, telephones, and electric utilities) have played a major role in generating growth for a very long time.

Sustained growth is certainly possible, but it is not likely to occur without the aid of outside driving forces, such as population growth and the relatively rapid growth of certain types of capital-intensive industries.

CHAPTER 11

Fluctuations in Income

In the last chapter we tried to show that the basic source of the secular growth of demand is the operation of a capital-adjustment process interacting with autonomous investment. For convenience we used an equilibrium approach. Now we have to ask why income fluctuates instead of growing steadily.

There are two basic approaches to the problem. Many writers try to explain fluctuations in income in terms of a stationary- or fixed-parameter system. In doing so, they either ignore autonomous investment or assume that it takes place in some regular way. Of course, no one supposes that the technical and behavior parameters governing income movements are really fixed. Those who explain the cycle in terms of a fixed-parameter system wish to show that a regular cycle would occur as a result of the system's basic characteristics. If they are correct, changes in parameters merely disturb the course of a cyclical process which would go on even if behavior parameters never changed.

The second approach is directly opposed to the first in that income fluctuations are explained in terms of changes in parameters resulting from changes in autonomous investment, speculation, changes in technique, and so on.

In the next section the major variants of the first group of theories are outlined, and it is shown that they appear to be inconsistent with some salient facts of economic history.

The argument given in section I appears to rule out the possibility that depressions prior to 1929 occurred because of a consistent tendency for capital to grow faster than income during the boom periods. We shall therefore assume throughout the rest of this chapter that we have to deal with systems which are capable of sustained growth.

In section II we consider a number of different types of exogenous or semiexogenous events which may cause fluctuations in income. It is shown that any combination of these events may cause a downward movement in income. The magnitude of the downswing is shown to depend in part on the nature of the original cause but also on the stability properties of the system. The influences of various aspects of the economic structure on the stability of the system are then discussed in detail.

In the next chapter the conclusions reached here are applied to two specific problems. It is first shown that ceilings are not likely to cause major depressions. We then list briefly some of the exogenous factors which have been important in causing depressions in the United States. Finally, we try to show why the depression of the 1930s was so much worse than its predecessors. It is argued that changes in the structure of the economy tended to reduce both the rate of growth of income and the stability of the system. Those tendencies were temporarily masked by some special features of the 1920s.

I. ENDOGENOUS CYCLE THEORIES

All the theories which explain business cycles in terms of the fundamental dynamic properties of the economic system have a common starting point. They all begin by supposing that the growth of income is produced by a fixed-parameter–capital-adjustment process of the type described in the last chapter. Since the income-generating process involves many lags and cumulants, and many interacting sectors, it is almost certain to produce cyclical movements in income unless the initial values of the variables are fairly close to the equilibrium values. It is therefore very easy to explain income movements by any of a large class of dynamic processes. The difficulty is that the cycles produced by such mechanisms cannot maintain a constant amplitude (if the equations of the system are linear). The cycles must either increase or decrease in amplitude as time passes. That is not in fact a very serious objection, since it is very unlikely that the true behavior equations really are linear.

Linear approximations may hold over a certain range of values of the variables, but that does not mean that the system is really linear over wide ranges. One can imagine that the dynamic process described in terms of linear approximations holding near the equilibrium values of the variables tends to produce explosive or antidamped cycles. If the equations of the system are actually nonlinear, the explosive tendencies may be checked when the cycle reaches a certain amplitude. After that, income will repeat a series of cycles of fixed maximum amplitude. Nonlinearities can be introduced into the system in many ways. The supply-of-funds equations underlying the investment theory used in the last chapter are certainly nonlinear. Linear approximations are convenient, but they cannot be taken seriously when linearity is important to the argument.

Moreover, a linear system producing antidamped cycles may be made to produce cycles of constant amplitude by introducing a limit on the rate of capital decumulation during the downswing.

Endogenous capital-adjustment processes (or generalized multiplier-accelerator systems) can produce cycles in several different ways. (1) A linear system whose largest root is imaginary will produce cycles. This corresponds to the case in which there is no stable position for the ratio of capital to income at which the rate of growth of capital equals the rate of growth of income. In such systems, income may grow for a time, given favorable initial conditions, but the rate of growth of capital will exceed the rate of growth of income. Eventually the rate of investment will fall so low that income begins to fall. No recovery can take place until capital has been decumulated. (2) A linear system which is capable of growing steadily may be prevented from doing so by ceilings on resources. The ceilings may cause the rate of growth of income to fall below the rate of growth of capital. As a result, investment may decline and cause a decline in income. (3) There may be nonlinearities in the behavior equations themselves. These will have the effects of causing capital to accumulate faster than income at some point in the same way as in the other cases. However, the time pattern of income movements will be different from the ones occurring in the linear cases.

The unconstrained linear models do not seem to fit the facts of the depressions before 1929. They imply that the rate of growth of capital exceeds the rate of growth of income for some time before income actually begins to fall. That must be so because no appreciable change in the ratio of capital to income can occur in a short period (unless we introduce outside shocks). The stock of capital rarely increases by more than about 4 per cent per year. If income is still increasing, the ratio of capital to income must increase by less than 4 per cent per year. It follows that the moment of absolute decline in income must be preceded by a long period during which profits are slowly falling and the rate of growth of income is slowly falling. That does not seem to square with the experience of the period since the Civil War.

Essentially the same argument applies to the theories based on ceilings. In addition, those theories are of dubious value because there is little evidence that resource limitations have played an important role at the end of most periods of prosperity.

Moreover, neither type of theory gives a very satisfactory account of the recovery process.

Let us consider the type of cycle produced by processes of the multiplier-accelerator type, however generalized. Suppose we have to deal with a system which contains no large exogenous expenditure component. Suppose income begins to fall, not relative to the equilibrium growth path but absolutely. What makes it turn up again? Once income starts down, it cannot recover unless capital decumulates, for, in all capital-adjustment processes, income turns down when capital accumulates faster than income. Income turns up only after capital has begun to grow more slowly than income. If income is growing negatively, capital must decumulate at a faster rate if income is to start to rise again. That is obviously true in the processes in which investment is made to depend on capital stock and income. But the acceleration principle (modified by the qualification that the principle does not hold when there is excess capacity) is simply another way of stating that investment depends on capital stock and income (except that it imposes severe limitations on the parameters).

If endogenous cycles require capital decumulation to explain the upswing from a depression, they cannot be used to explain any observed upswings except the upswing of the 1930s, for gross investment never fell by more than about 40 per cent in any depression before 1929. Net investment must certainly have remained positive in the aggregate.

In a multisector system it is possible for income to change its direction of movement if the stock of one type of capital is decreased while the stock of another type increases by a greater amount. In that case recovery occurs without decumulation of capital in the aggregate. What is required, of course, is that the increase in investment (*cet. par.*) caused by a unit decrease in the stock of one type of capital be greater than the reduction in investment caused by a unit increase in the stock of the other type. Inventory investment, for example, is much more sensitive to changes in the stock of inventories than most types of fixed capital. But inventories aside, it seems doubtful whether there was capital decumulation (of any magnitude) in any major sector during the depressions before 1929.

The depressions of 1907 and 1920 were so short that no appreciable amount of decumulation could have taken place in any sector. It is hard to see how an endogenous process with fixed parameters can be used to explain those depressions. If those depressions are ruled out, we have a period of thirty-two years without a cycle produced by an endogenous process. That is not much of an argument for a regular-cycle theory. It should be noted that these remarks apply to ceiling theories as well as to other endogenous theories. The situation is less clear in the cases of the depressions of 1873, 1883, and 1893. We do not have enough annual data on capital formation to know whether any type of capital decumulated in those depressions. But since there was no aggregate decumulation, the burden of proof is on those who wish to argue that decumulation in one sector more than offsets accumulation in other sectors.

It may be argued that recovery can occur without capital decumulation if there is autonomous investment. That is true if autonomous investment is simply a lump of expenditure which has no adverse effect on other expenditure. But we have already seen that that is not the case. It was shown in the last chapter that autonomous investment has the same effect as any other factor which shortens the lag between income and expenditure. If the amount of autonomous investment is constant, the adverse effect of earlier investments will cancel the favorable effect of the current ones without assisting the recovery process. A recovery could be produced if the amount of autonomous investment grew steadily, but it seems very artificial to rest a regular-cycle theory on the notion of regularly increasing autonomous investment.

A depression is likely to retard rather than encourage autonomous investment. It somehow seems unlikely that autonomous investment will rise during a depression unless we suppose that some outside shock had previously caused it to decline. In that case, of course, the depression would have been caused by the shock rather than by the operation of an endogenous system.

Models employing nonlinearities in the behavior equations can produce a wide variety of patterns of income movement and are therefore much less restrictive than linear models. For example, a model employing a rigid limit on the ratio of debt to profits could be used to produce a cycle in which the rate of growth of income would suddenly change from positive to negative. It could be supposed that debt has no adverse effect on investment until a certain critical value is reached. If we begin from a position in which the debt ratio is below the critical figure, investment and borrowing might take place and income would rise. With the right behavior functions, debt might be rising faster than profits, so that at some point the critical ratio would be reached. Investment would then fall suddenly and income would begin to fall. However, a sudden change in the rate of growth will occur only if all firms (or a large proportion of them) reach their debt limits at the same time. That somehow seems implausible.

It is possible to construct a variety of nonlinear models which produce cycles. But thus far no plausible model has been offered which explains cycles of the kind we have observed.

The arguments given above seem to rule out endogenous processes, whether unconstrained or equipped with ceilings and buffers, as explanations of observed major cycles.

Observed changes in income and capital formation have not been like the ones associated with endogenous-cycle models of the capital-adjustment type. It is very probable that any system of equations describing a capitalist economy does have some cyclical roots. One can reasonably ask why those roots do not produce any cycles. In fact, of course, they undoubtedly do produce many cycles in the sales and capital formation of particular industries. They may also produce some minor cycles in the movements of aggregate income. The real problem arises in connection with major cycles. If we do not get endogenous cycles, it must be because of the properties of the endogenous system.

Consider a capital-adjustment process of the type we have been discussing. Suppose its largest root is real and greater than 1. Then the system is always stable from above. That is, if the variables take on values which produce a rate of growth higher than the equilibrium rate, the variables tend to move in such a way as to reduce the rate of growth to the equilibrium value. Suppose that the system also has considerable stability in the downward direction. That is, (1) the rate of growth tends to converge to the equilibrium from any average value of $\frac{k_i}{y}$ which does not exceed the equilibrium by more than, say, 15 per cent (neglecting the possibility that the deviations of individual k_i's from their equilibrium values are selected in such a way as to maximize their adverse effect); (2) if investment is reduced arbitrarily by a substantial amount, income declines at a fairly slow rate.

It is shown below that there is reason to believe that the first condition has usually been satisfied for the American economy. The relatively slow downward movement of income in the 1870s and 1890s indicates that the second condition was satisfied. (The system was not really stationary during the period in question, but the exogenous elements in the downswing in the 1870s and 1890s did not contribute much to the recovery.)

If the two conditions given above are satisfied and if the system starts out with initial conditions in the right range, the cyclical roots will never have a chance to come into play unless the parameters of the system are disturbed and remain abnormal for a rather long time. Fluctuations in income can occur, of course, but neither the downswing nor the upswing can be accounted for by the cyclical characteristics of the endogenous system.

We must look elsewhere for the causes of fluctuations in income. It is argued below that, in the past, income fluctuation has been due to the operation of a combination of exogenous factors working on the system. These include variations in autonomous investment, the adverse effects of speculative booms associated with the development of growing industries, the stock market, or with wage-price spirals. Exogenous changes in the foreign balance and in monetary conditions have also played a role in producing fluctuations.

We shall also try to show that labor shortages, monetary shortages, and capital-goods shortages are much more likely to produce minor cycles than major cycles.

II. EXOGENOUS CAUSES OF FLUCTUATIONS

A stationary (fixed-parameter) economic system can generate movement in income and the other economic variables, but those variables can never follow the path corresponding to a given stationary system for very long. The system is constantly being disturbed by technical changes, changes in the rate of population growth, development of new areas, changes in financial institutions, changes in competitive structure, and so on. It is only by imagining that those changes do not occur at all or that they occur in some steady way that we can conceive of steady growth. It is obvious that steady growth of income is most unlikely in a world whose techniques and institutions are constantly evolving. In addition, we know that capital-adjustment processes exhibit a certain tendency to produce cumulative processes. The problem is not so much to explain the fact that fluctuations occur as to find out what limits their magnitude and why some of them are more violent than others.

We can first consider the major classes of disturbance to which the system is subject. Then we can attempt to see what factors determine the extent of the system's reaction to those disturbances.

The disturbances fall into a number of major classes.

1. Fluctuations in autonomous investment may occur for a number of reasons. Such fluctuations may, of course, occur simply because of variations in the rate at which innovations occur or in the type of innovation. It seems obvious that the importance of autonomous investment varies from period to period. However, one would expect that, if only technical factors were involved, autonomous investment would increase and decrease gradually over relatively long periods. But the exploitation of technical developments may be hampered during wars so that backlogs of autonomous investment are built up. In that case autonomous investment would proceed rapidly after the war and then taper off to a lower level. One can also imagine that the unfavorable financial conditions of a depression hinder the exploitation of autonomous investment. A backlog of unexplored possibilities is then built up. Those possibilities eventually contribute to recovery and a subsequent boom, but again the backlog will eventually be worked off.

In any case, there are fluctuations in the rate of autonomous investment. During the upswing of such a fluctuation income will grow rapidly. As the rate of increase of autonomous investment falls, the rate of growth of income tends to fall. When autonomous investment reaches a peak it is making no net contribution to the income stream (except to the extent that the growing industries are more capital-intensive than older ones). When autonomous investment begins to decline, the rate of growth of income is reduced by the adverse effects of earlier autonomous investments. However, it may still continue to grow at a rate above its equilibrium rate because of the lagged responses of consumption and some types of investment to the earlier high rate of growth.

Suppose the system would have an equilibrium rate of growth in the absence of autonomous investment. The lagged effects of the high rate of growth during the upswing of autonomous investment tend to keep the actual rate of growth above its equilibrium value for a time, but increasing sales of the new products introduced during that period have the opposite effect. Note that both effects will wear off after a time. There is no a priori way of telling which effect will be the larger. If the negative effects of the earlier innovations are sufficiently large, income will turn down for a time. However, it will have a *tendency* to recover when the delayed effects of the innovations have been absorbed. It will be shown later that the tendency for income to recover may be offset by other consequences of the decline in income.

If the system has no equilibrium rate of growth (without increasing autonomous investment), income is bound to turn down after a time. It cannot begin to rise again until autonomous investment rises.

There are, of course, no pure cases in which a fluctuation in income is produced solely by the movement of autonomous investment, but such movements may contribute to the violence of the swings to which many other causes also contribute.

2. Closely allied to autonomous investment is speculative investment. We can define speculative investment as investment which is based on an unwarranted estimate of the future movements of the variables determining the profitability of the investment. That is a somewhat vague definition, but there are situations to which it clearly applies.

A long period of prosperity tends to produce speculative booms of three different types:

a. The profit prospects of a particular industry which is growing rapidly may be exaggerated. The ratio of capital stock to demand for the industry may be raised for a time. When the boom collapses, because the profits do not materialize or for some other reason, investment will fall sharply. It will fall not only because the industry has too much capital but because the growth prospects of the industry are underestimated for a time. If the industry grows into its capital, and investors take a more reasonable view of its prospects, the recovery of the industry will help to speed up the rate of growth of income.

b. Long periods of prosperity favor security-market speculation.

The high price of equities encourages investment. The stock-market boom may be brought to an end by some random shock or because the rate of increase of stock prices is checked by limitations on the supply of funds. During the ensuing panic, investment will fall because of the change in expectations and the difficulty of obtaining funds. If nothing else has happened, investment will tend to recover after a time, though not to its boom level.

c. If income grows very rapidly for a time, labor may become scarce. When the forces behind investment demand are particularly strong, e.g., in a postwar boom, investment may continue at very high levels and generate a rapid increase in money demand, even though the real rate of growth has slowed down. The persisting labor shortage may lead to a wage-price spiral. High levels of investment may then be based on expectations of rising prices. Such a boom is bound to end for one reason or another. When it does, investment is bound to decline.

3. Shortages of labor may lead to speculative booms, and shortages of money may bring them to an end. But it is possible to have a shortage of labor without bringing on a speculative boom, and it is possible to have tight money when there is no stock-market boom. Shortages of labor, money, or capital goods restrict the growth of investment and income directly. As a result, capital may accumulate faster than income and cause a decline in investment demand.

4. The parameters of the system are subject to a variety of random changes due to technical changes, the effect of political events on expectations, changes in demand for exports, or arbitrary changes in monetary and fiscal policy. For the most part the effects of these changes will cancel out, but occasionally a number of them may operate in the same direction, or they may reinforce the effect of some of the other factors to be discussed later.

5. Changes in taste or technique may produce systematic trends in the parameters of the system which tend to slow down the rate of growth of income.

Given that an equilibrium rate of growth exists, it is possible that a major depression might arise in the process of transition from one equilibrium to another. Any persistent parameter change will, of course, cause the equilibrium rate of growth of income to change in the manner described in the last chapter. However, the system will not jump from one equilibrium position to another. It has to go through a transition process which may last for a considerable time. While the transition from one equilibrium rate of growth to another is going on, the actual rate of growth need not lie between the old and the new equilibrium rates. If the average capital intensity of the system begins to decrease, the equilibrium values of $\frac{k}{y}$ and r_y will fall. The initial value of $\frac{k}{y}$ will be above the equilibrium value, and the initial value of r_y will lie not only below the old equilibrium value but below the new one as well. As a result, income may tend to fall while capital is being decumulated. The extent of the decline in income which has to take place before the system can reach the new equilibrium position will depend on the speed with which income falls during the adjustment process.

It is conceivable that a depression of considerable magnitude could occur in the process of transition from one equilibrium position to another. In fact there do not appear to be any examples of a depression of that type. The reason is that most persistent parameter changes take place fairly slowly, so that the adjustment process never leads the system far away from its equilibrium for the current set of parameters.

In all the cases cited we have an initial change in the parameters of the system which tends to reduce the rate of growth of income (or make it negative). And in all cases (except that of trendlike changes in parameters) there is a tendency for the initial change to reverse itself. The effectiveness of the latter tendency, however, depends on the nature of the income and price movements which take place as a result of the initial downward shift in investment demand.

III. CONDITIONS FOR RECOVERY

It can be said, therefore, that every shock which produces a depression tends ultimately to produce conditions which are favorable to recovery. But whether a recovery actually occurs (without a long period of capital decumulation) depends on what happens during the downswing. The nature of the factors determining whether the recovery takes place can be illustrated by reference to depressions caused by financial crises and speculative booms in particular industries.

In a depression brought about by a financial crisis (however caused) investment demand will shift downward because (1) equity prices fall, (2) underwriters are reluctant to undertake new issues in view of the uncertainty of the situation, (3) banks and other institutional lenders try to reduce rather than extend their commitments in order to gain liquidity, (4) firms which normally extend trade credit become unwilling to do so because of the danger of default, (5) expectations deteriorate.

The resulting reduction in investment causes income to decline, and prices will also decline. After a time the reduction in money demand will in itself tend to eliminate the financial crisis. The reduction in transactions demand makes for easy money—bond yields tend to fall, trade credit becomes easier, and good fixed-interest securities can once more be floated. The investment-demand function will tend to shift upward again, and, unless income has declined a great deal relative to capital stock, investment and income will begin to rise again. However, that tendency can be frustrated if the rate of decrease of income and prices and the rate of deterioration of expectations have become very great before the initial liquidity crisis is over. In that case a series of failures of business firms and financial institutions will increase the demand for liquidity and, together with the deterioration in expectations, will cause further downward shifts in investment demand. Income will then have to fall a long way before a real recovery can begin.

It is clear that the amplitude and duration of the depression resulting from a financial crisis will depend on the speed with which income and prices move downward as a result of the initial downward shift in investment and on how long it takes to eliminate the initial crisis.

The same thing is true of depressions brought on by speculative booms in particular industries. In that case, the prospects of the industry are overestimated during the boom, and the industry acquires more capital than would be justified by a more level-headed estimate of the industry's growth prospects. When the boom breaks, the industry's prospects are underestimated for a time. If the decline in aggregate demand proceeds slowly, the industry will eventually grow into its excessive capital stock (because its demand is growing relative to aggregate demand). When its profits rise, the industry's prospects will again be reestimated, and its investment will rise. The resulting tendency, which is for income to increase, will be reinforced if the industry which has had the boom is more capital-intensive than most other industries. In that case the reduction in the capital-demand ratio of the industry in question which results from growth in its share of aggregate demand will increase investment. The increase will more than exceed the reduction in investment in the industries losing from the shift in demand.

The tendency for the investment-demand functions of the growing industry to shift upward will be frustrated if income falls too rapidly. If aggregate income moves down fast enough, the demand for the growing industry will decline in spite of its increasing share of aggregate demand, and there will be no upward revision of estimates of its future prospects. Eventually, of course, aggregate demand will stop falling, but it is clear that, as before, much depends on the speed of the cumulative downward movement in aggregate income.

Similarly, after a speculative boom brought on by a wage-price spiral, inventories will be liquidated, and the excessive pessimism of the panic will disappear with time. But that will be true only if income has not declined rapidly during the panic.

Finally, shortages of labor, capital goods, and credit do not usually last very long, but much depends on what happens to income and investment while shortages are holding down the growth of income.

In each of the cases cited the following elements are present: (1) an adverse shock which causes a reduction in the rate of investment; (2) a fall in the rates of growth of both income and capital stock (the reduction in the rate of growth of income will nearly always exceed the reduction in the rate of growth of capital stock); (3) an increase in the ratio of capital stock to income; (4) further reduction in both saving and investment as fractions of lagged income in response to (3); (5) continued decline in income at a rate determined by the initial reduction in the rate of investment plus the secondary reduction in investment resulting from decline in income offset by the induced reduction in the savings ratio; (6) induced changes in behavior parameters as a result of induced price changes, changes in expectations, and so on; (7) effect of (6) on rates of saving and investment and on rate of decrease of income; (8) changes in behavior parameters due to elimination of original adverse shock; (9) the resulting tendency for investment to increase, which may be sufficient to cause income to begin to increase; (10) if the recovery tendency does not cause income to increase at the time when it takes place, recovery cannot occur until there has been sufficient capital decumulation to reduce the ratio of capital to income.

IV. AN APPROACH TO THE ANALYSIS OF RECOVERY PROCESSES

So far we have argued that (1) adverse shocks can cause depressions and (2) there is some tendency for the shocks to reverse themselves and create conditions favorable to recovery. Now we want to examine the conditions under which an upward shift in investment can produce a recovery. We shall first try to show that the effectiveness of an upward shift in investment after a period of depression depends on (*a*) the relative sensitivity of savings and investment to a given change in the ratio of capital stock to income, (*b*) the extent of the increase in the capital-output ratio during the downswing. We shall then show that the extent of the change in the capital-output ratio during the downswing also depends on the relative sensitivity of savings and investment to changes in capital-output ratios. On both counts, then, the chance of recovery depends on the responses of savings and investment to capital-output ratio.

The argument can be summarized as follows:

In section A we assume that a shock has caused a downward shift in the investment function and a downward movement of income. We then suppose that the investment function returns to its original position. We suppose that in the interim the ratio of capital to income has risen. It is then shown that income will begin to rise at that point if the decline in the savings ratio induced by the rise in the capital-output ratio exceeds the induced decline in the ratio of investment to income.

In section B it is argued that the induced decline in the ratio of investment to income increases relative to the induced decline in the savings ratio as the increase in the capital-output ratio grows larger.

It follows that if the adverse shock has a given duration, the chance of recovery depends in part on the speed with which income moves down under the impact of the shock.

In section C it is shown that the speed of movement of income also depends in part on the relative sensitivity of savings and investment to the change in capital-output ratio.

A. Effect of Given Change in Capital-output Ratio

It was said above that the relative sensitivity of saving and investment to changes in capital stock and income plays an important role in determining the amplitude and duration of a depression.

To keep the argument as simple as possible, suppose that at the onset of the depression the system is in equilibrium. The rate of growth of income equals the rate of growth of capital stock. Some adverse shock causes a downward shift in the relation between investment, capital stock, and income. For simplicity, suppose that after a time the effect of the shock is completely reversed. The pre-depression relation between investment capital stock and income holds once more. Meanwhile, however, income will have fallen and capital stock will have risen. The ratio of capital to income will be higher than before the depression. As a result the rate of investment as a fraction of income will also be lower than before the depression. The rate of saving as a fraction of income will also be lower than before the depression. The rate of change of income will be reduced by the difference between the fall in the ratio of investment to income and the fall in the ratio of savings to income. The rate of increase of capital stock will also fall because (*a*) the ratio of investment to income has fallen, (*b*) the ratio of capital stock to income has risen.

The necessary condition for a recovery of income (after the reversal of the shift in the investment demand) is that the rate of change of income should be equal to or greater than the rate of change of capital stock. That condition will be satisfied if the induced decline in the savings ratio is large relative to the induced decline in the ratio of investment to income. The savings ratio and investment ratio are both functions of lagged income. (Also, the ratios are ratios of planned saving and investment to lagged income.)

The meaning of the condition can be illustrated by an arithmetical example. Suppose that before the depression the ratio of capital to annual income is 2.5, net investment in each period is 10 per cent of the previous period's income, and consumption in each period is 92 per cent of the previous period's income. There are two income periods per year. Capital stock is rising at 2 per cent per period, or 4 per cent per year. The system is in equilibrium. Now let the investment-demand function shift downward and then, after a time, return to its original position. Meanwhile income, has been falling and capital has been accumulating. Suppose that the ratio of capital stock to income rises to 3 before investment demand recovers. Let us suppose that as a result the ratio of net investment to lagged income falls to 6 per cent. Capital stock is increasing at 2 per cent per year. If income is to recover, it must rise at over 2 per cent per year, or 1 per cent per period. If the savings ratio falls to less than 5 per cent, that condition will be satisfied.

The figures used are more or less arbitrary, but they indicate the principle involved. The principle involved in this argument is illustrated by the stability range shown in Figure 9. The stability-range concept is too simple because it does not take account of lags, but the basic idea is valid even in a complicated system with many lags.[1]

[1] In the argument given above, it was assumed that the investment function returned after a time to its original position. However, that is not necessary in order

Basically, then, the chance for an automatic recovery from the initial shock depends on the relative sensitivity of investment and saving to changes in capital-output ratios.

B. Effect of Size of Change in Capital-output Ratio

So far we have shown that the chance of a recovery (without a long period of capital decumulation) depends on the relative responses of saving and investment to the changes in capital and income which take place as income declines. It is also true in general that the ratio of reduction in saving to reduction in investment is

---

to produce a recovery. In some cases it is possible for the system to reach a new positive equilibrium rate of growth on the basis of a lower investment function than the initial one. Suppose that investment as a function of capital stock and income is reduced. That reduces the rate of growth of income and the rate of growth of capital stock. The rate of growth of income will fall by more percentage points than the rate of growth of capital stock. (The rate of growth of income falls by the same number of percentage points as the decline in investment as a percentage of income. The rate of growth of capital stock falls by the same figure *divided* by the initial ratio of capital stock to income, which always exceeds 1.)

(Saving is presumably not affected by the shock except in special cases.) If the system was in equilibrium to begin with, the initial rates of growth of capital and income were equal. After the shock the rate of growth of capital exceeds the rate of growth of income, and the ratio of capital to income tends to increase. As it does so, the rate of investment (as a fraction of income) tends to fall, and with it the rate of growth of income. The rate of growth of capital also falls as a result of the increase in capital-output ratios. As we have seen, the rate of growth of capital will fall faster than the rate of growth of income if the propensity to save is sufficiently sensitive to increase in $\frac{k}{y}$. For a time, then, $\frac{k}{y}$ will continue to rise, and the rate of growth of income will rise relative to the rate of growth of capital, thus reducing the gap opened up by the initial shock. Suppose that the initial shock is small or that the sensitivity of saving to capital-output ratio is large relative to the sensitivity of investment to those ratios. Then after a time the difference between the induced fall in the rate of growth of capital and the induced fall in the rate of growth of income will equal the difference in the rates caused by the initial shock. The system will have reached a new equilibrium. If that does not occur after $\frac{k}{y}$ has increased by a certain amount, it will never occur. For once $\frac{k}{y}$ becomes sufficiently large, further increases reduce the rate of growth of income by more than the rate of growth of capital, and the gap between the two rates begins to widen again. Recovery then depends on the occurrence of an upward shift in the investment function.

likely to be greater for small increases in $\frac{k}{y}$ than for large ones. That is so because of the influence of lags. If income has been rising fairly rapidly (as is usually the case in the years before a major depression), consumption, dividends, and the rate of residential construction tend to be below their stationary values for the existing capital stock and income (that is, they are below the values they would reach if the income and capital stock remained fixed for some time). When income declines by a small amount, dividends, consumption, and residential construction fall relatively little (indeed they may even rise) so that the savings rate falls by a very large proportion of the reduction in income, and investment by a much smaller proportion. As the reduction in income grows larger the influence of the lags becomes relatively smaller so that the ratio of reduction in savings to reduction in income grows larger as the size of the reduction in income grows larger.

C. Speed of Movement of Income

In any given length of time the extent of the reduction in income depends on the speed with which income declines. We shall show that the speed with which income falls depends in part on the relative sensitivity of saving and investment to changes in the ratio of capital stock to income.

First, however, it should be noted that the speed with which income declines has a direct effect on the investment function. The speed with which income falls influences price and income expectations, which in turn influence the position of the investment function.

An example will indicate the nature of the link between the speed of movement of income and the responses of saving and investment to changes in capital stock and income.

Suppose that initially the ratio of capital stock to the previous period's income is 2.5. The income period is a half year, and flow variables are measured at annual rates. Suppose consumption each period is 92 per cent of the previous income and net investment is 10 per cent of the previous period's income. Then income rises by 2 per cent per period, or 4 per cent per year. Capital stock is rising by 10 per cent of income or 4 per cent of capital stock and the system is in equilibrium. Now let net investment be arbitrarily reduced by 30 per cent of itself to 7 per cent of lagged income. Income will now fall at 2 per cent per year, and capital stock will rise by slightly less than 3 per cent per year. The reduction in income and the rise in capital stock will reduce both saving and investment as fractions of lagged income. If only investment fell, the rate of decrease of income would rise rapidly.

But in fact, of course, saving will fall as well. If both ratios fell to the same extent, income would continue to fall at 2 per cent per year. Capital stock would continue to rise, but at a progressively slower rate. If the savings ratio declined faster than the investment ratio, the decline in income would be retarded, and conversely. Just as a matter of arithmetic, then, the rate at which income falls depends on the relative response of saving and investment to changes in capital and income. Those responses will, of course, change as the downward movement proceeds.

Our task now is to survey a number of aspects of the economic structure which influence the movements of saving and investment.

V. SAVING, INVESTMENT, AND THE SPEED OF MOVEMENT OF INCOME

In the last section we argued that the speed of movement of income in response to a downward shift in investment demand plays a crucial role in determining the amplitude and duration of a depression. We then argued that the speed of movement of income depends on the relative responses of saving and investment (relative to income) to changes in income relative to capital stock.

The factors determining the movements of saving and investment in response to changes in income fall into three groups: (1) the initial size of various responses to changes in income and capital stock when lags are disregarded, (2) the influence of various types of lags, (3) changes in the structure of the system induced by the fall in income. We shall take those factors up in turn. In each case we have to consider the factors affecting both saving and investment.

A. Unlagged Responses of Saving and Investment

We have first to consider the relation of the savings ratio to changes in income. If we neglect lags, we can assume that the ratio of personal consumption to disposable income is fixed. We can concentrate on business and government saving.

1. *Saving*

The sensitivity of the savings ratio to changes in income relative to capital stock will depend very largely on the response of the ratio of profits before depreciation to changes in income (capital stock given).

The sensitivity of gross-profit margins with respect to income (capital stock given) depends on two factors: (*a*) the sensitivity of prices to changes in demand, (*b*) the shapes of cost curves.

It is fairly obvious that profits as a percentage of income will fall if prices fall relative to wages. The degree of short-run price flexibility in the system depends on various aspects of the competitive structure. These include not only the numbers of firms in various industries but also a great variety of factors, such as the nature of the opportunities for secret concessions to dealers in the form of discounts, advertising allowances, and so on.

Even if prices do not fall relative to wages, gross-profit margins are likely to fall with reductions in volume. At the top of a boom many firms will be in the rising range of marginal costs because they are working overtime, using extra shifts, or using obsolete equipment. A reduction in output is therefore likely to reduce direct labor cost per unit of output. But a very considerable part of the cash costs, i.e., costs other than depreciation charges, of industrial firms is fixed or at least relatively fixed. These fixed cash costs include interest, property taxes, insurance, pensions, a part of advertising expense, and a large part of the cost of management, supervision, and research. Some of those items can be reduced under sufficient pressure or a cut in volume which results in closing down plants. But for short-run variations in output they are likely to remain fixed. The fixed cash cost per unit will rise as output falls and cancel any fall in direct labor cost per unit.

The extent of the reduction in gross business savings will depend on the extent of price reductions and the relative importance of fixed cash costs.

Government revenues from the corporate income tax will decline in proportion to the decline in net corporate profits. Because of the leverage of fixed costs the reduction in net corporate profits will be more than proportional to the reduction in income. Government revenue will decline by a greater percentage than personal income because of the form of the personal income tax.

If government expenditures remain fixed or rise (because of rising transfer payments), government saving will fall rapidly with a reduction in income.

2. *Business Investment*

The response of business investment to changes in income relative to capital stock will depend, of course, on the position and movements of the marginal-efficiency-of-investment schedules of firms and their supply-of-funds schedules.

A reduction in income will cause marginal-efficiency schedules to shift down to the left and will cause supply-of-funds schedules to shift upward to the left.

The extent of the shift in supply schedules will depend on the movement of profits (discussed above) and on the initial debt position. If firms have a relatively large amount of debt outstanding, the leftward shift of the supply schedule will be relatively great because debt remains the same while profits fall. Other things equal, then, investment will be less stable the greater the ratio of debt to profits at the start of the downswing.

The amount of downward shift in investment demand for a given reduction in income depends primarily on technical factors (though it also depends to some extent on the competitive structure). The significance of the downward shift, however, depends on the elasticity of the marginal-efficiency schedule and the initial position of the schedule relative to the supply-of-funds schedule.

If the marginal-efficiency schedule is highly elastic and (before the downswing) intersects the supply schedule in the inelastic range, a downward shift in the schedule will make little difference. The reduction in investment resulting from a fall in income will depend primarily on the movement of the supply schedule.

If, on the other hand, the marginal-efficiency schedule is less elastic or at a lower level, investment will be depressed by the downward movement in investment demand as well as by the fall in profits.

3. *Housing*

The stability of investment in housing will depend on the income elasticity of demand for housing and on the rate of population growth. If there is no population growth, the demand for housing will decrease by an amount (per year) equal to the rate of decrease of income multiplied by income elasticity of demand for housing.

If there is population growth (or rather growth in number of households), demand for housing will fall at a slower rate. The reduction in the rate of decrease of housing demand will equal the percentage rate of increase in households times 1 minus the income elasticity of demand for housing. (The formula for computing the increase in demand for housing is: Percentage rate of increase in demand equals percentage rate of increase in income per household times income elasticity of demand plus percentage rate of increase in number of households.)

Of course, the actual stability of the rate of construction will depend on the size of the backlog of demand built up by credit rationing.

VI. EFFECT OF LAGS ON SPEED OF MOVEMENT OF INCOME

It is fairly obvious that the speed of movement of income under the impact of an adverse shock will depend very largely on the lag structure of the system.

In the last chapter we took the lags into account by introducing the rate of growth of income into the behavior equations. Since we were concerned only with equilibrium, it was necessary to consider only a single rate of growth. When we turn to the adjustment of the system to changes in parameters, it is necessary to consider changing rates of growth, with the result that the lags enter the system in a more complex way. The reactions of the variables to changes in income will depend on the past history of income. For example, if income has been rising at a high rate for a number of years and then begins to fall, consumption will fall more slowly than if income had been constant in the years before the downswing began. Since the relevant part of the history of the system is constantly changing, the ratios of consumption to income, dividends to profits, and housing demand to income are constantly changing even for given values of the basic parameters.

This raises the question whether it would not be better to put the argument in terms of a formal sequence analysis more suitable to a dynamic problem than the scheme used for the equilibrium analysis in the last section. Even though it involves some clumsiness of statement and some loss of accuracy, it seems useful to retain the formal scheme used in the equilibrium analysis and allow the parameters of the system to change as the history of past rates of growth changes. The advantage in doing that lies in the possibility of retaining an intuitive grasp of what is going on.

If we keep the consumption equation in the form $C_t = Ay_{t-1}$, we have to regard A as a function of the pattern of past movements of income and ask how it changes as the pattern of income movements changes. The answer is fairly simple. The original equation was

$$C_t = Py_{t-1} + qC_{t-1}$$

which is equivalent to

$$P(y_{t-1} + qy_{t-2} + q^2y_{t-3} + \cdots$$
$$= Py_{t-1}\left(1 + q\frac{y_{t-2}}{y_{t-1}} + q^2\frac{y_{t-3}}{y_{t-1}} + \cdots\right.$$

$\dfrac{C_t}{y_{t-1}}$ increases as the ratios of the past values of income to last period's income rise (the importance of the more recent incomes being, of course, much greater than that of earlier ones). $\dfrac{C_t}{y_{t-1}}$ will be largest when income has been falling rapidly for some time and smallest when income has been rising rapidly for some time. If income has been rising and then begins to fall, $\frac{C_t}{y_{t-1}}$ will start off at a relatively low value and then begin to rise. When the decline in income takes place at a constant rate, $\frac{C_t}{y_{t-1}}$ will rise at a decreasing rate and converge to a constant value. When income begins to rise again, the process will reverse itself. In a short cycle when income changes its direction of movement twice in a short period, the value of $\frac{C_{t-1}}{y_{t-1}}$ will move only slightly because the influence of the lagged values of income will cancel out.

The r_y curve will move up and down with A, since A appears as an additive term in the equation for r_y. If some parameter change which reduces investment takes place, income may decline. With a fixed r_y curve the initial decline in income might raise $\frac{k}{y}$ and cause a further decline in r_y. That tendency will be offset by the upward shift in the r_y curve which results from the operation of the consumption lag.

The effects of the lag in the response of consumption to changes in income will be reinforced by population growth. When population is growing, per capita income falls by the sum of the rate of decrease of aggregate income and the rate of growth of population. The ratio of consumption to income rises as the ratio of current income to past incomes falls. Therefore the larger the rate of growth of population the larger the average propensity to consume.

It seems clear that exactly the same argument applies to the movements of dividends and entrepreneurial withdrawals in relation to profits.

The housing problem is slightly more complicated. The lagged reaction of housing demand to income change has essentially the same effect on the r_y curve, i.e., on the relation of growth of income to the capital-income ratio, as the lagged reactions of consumption to disposable income and of dividends to profits. However, to the extent that the operation of the lag makes r_y shift upward when income begins to decline, it also makes the rate of growth of capital stock greater than it otherwise would be. Since the ratio of value of capital stock in housing to income is of the order of magnitude of 1, the effects cancel out for housing. If, as a result of the lag, the ratio of housing construction to income rises by .01, the ratio of new construction to housing stock is also increased by about .01. $\frac{k}{y}$ therefore moves in essentially the same way as it would have moved without the lag. But since the increase in construction influences only the capital stock in the housing industry, $\frac{k}{y}$ is reduced for all other industries. The lagged reaction of housing demand to income does, therefore, tend to slow down the cumulative downward process in essentially the same way as the lags in consumption and dividends. By the same token it acts as a drag on recovery from a long downswing. Since the lag in housing is somewhat longer than in the other cases, its effects may be felt for a considerable time after the recovery.

The lagged response of housing demand may be reinforced by the lagged response of the construction industry itself. That consideration is not of great importance after a rise in income has gone on for some time. But if there is a housing shortage as a result of a war or a period of very rapid increase in investment, the slow reactions of the construction industry on the upswing will reduce the secondary effects of the rise in other investment. They will also have a very marked stabilizing effect if the investment boom is interrupted.

A. Age of Capital Stock

We know, or at least have reason to believe, that the rate of investment depends on the age distribution of the capital stock as well as on its average age. As Domar and Eisner have shown, the device of measuring the net capital stock by applying straight-line depreciation is not an adequate way of gauging the effect of capital stock on the marginal efficiency of capital. That is so because the probability that it will pay to take any piece of equipment out of service is low for a number of years after its purchase and then rises rapidly. A piece of equipment five years old, therefore, has just about as much of an effect on the marginal efficiency of investment as a piece of new equipment, though it may be shown on the books as worth only half as much. If the age distribution were rectangular, that would make no difference. If some type of equipment had a ten-year service life, 10 per cent of the equipment would reach the end of its life each year, and straight-line depreciation would subtract 10 per cent from the depreciated capital stock.

When the capital stock is growing, however, the age distribution is skewed toward the younger ages. Less than 10 per cent of the capital stock reaches replacement age in each year. As we have already seen, this consideration acts as a drag on the rate of growth of income in the same way as the consumption and dividend lags. And just as in the case of the other lags, the replacement lag acts as a stabilizing factor in the downswing. When investment declines and the rate of growth of measured capital stock decreases, the proportion of the measured capital stock actually due for replacement (if in use) rises, and therefore the rise in the ratio of effective capital stock to income is less than the rise in measured capital stock. Of course the lag operates in the opposite direction when the rate of increase of income rises.

VII. INDUCED CHANGES IN STRUCTURE

A. Expectations

So far we have been dealing with factors which tend to retard changes in the rate of growth of income in either the upward or the downward direction. We now have to consider some factors which tend to work in the opposite direction. The first of these is expectations. We are now concerned with expectations about future movements of aggregate demand. Expectations of future movements of demand may change for many reasons, but at present we can confine ourselves to the changes in expectations which are due to actual movements of demand. In general, it seems reasonable to suppose that expectations are not very sensitive to small changes in the rate of growth when it is positive and is less than, say, 4 per cent or so per year. When rates of growth are of that magnitude, most firms will not be prepared to project growth into the future in making investment decisions. But when growth rates become much higher, some tendency to project them into the future is likely to arise. That will, of course, stimulate investment and may actually make the growth rate rise. Since high rates of growth of demand are often connected with labor shortages and wage-price spirals, they are discussed in Chapter 14. For the moment let us consider the other side of the picture, when the growth rate becomes negative.

Whenever income falls, even by a small amount, investors become wary. Even if they do not change their best estimates of future-demand movements, their uncertainty about the future certainly increases. It is obvious that an increase in uncertainty or a reduction in modal estimates of future demand will reduce the marginal efficiency of investment. It also seems likely that the reduction in marginal efficiency of capital will be larger the greater the rate of decrease of income. Such a reduction, of course, tends to reduce actual investment and to speed up the rate of decrease in income. The reaction of expectations to negative growth rates therefore tends to speed up cumulative processes in the downward direction.

Moreover, the strength of the adverse effect of changes in expectations tends to become greater as the cumulative decline in income from the peak of the boom becomes greater.

Small decreases in income are not infrequent, and they can often be attributed to special causes of a temporary nature. They may be fairly readily explained away and not have very much effect. But the greater the cumulative downward movement becomes, the more uncertain investors will be as to how much farther down income will go before it stops falling. At the same time, changes in the marginal efficiency of investment may have less effect in the early stages of a downswing than later on. At the top of a boom many firms will have a large backlog of investment projects which would be worthwhile carrying out even if income were somewhat below its present levels. The projects will be carried out as funds become available from internal sources. In that case a moderate deterioration of expectations has little effect. After income has fallen some distance, the stock of projects which would still be worthwhile even if income fell farther will have become much smaller, and a deterioration in expectations will have a much greater effect. The vicious spiral of interactions between expectations and negative rates of change of income therefore has some tendency to gain momentum as the downswing proceeds.

B. Prices and Price Expectations

If income declines to any appreciable extent, prices and wages are likely to fall. The prices of agricultural and other competitively produced raw materials will almost certainly fall as aggregate demand falls. As profits fall and competition becomes more severe, managements attempt to reduce wages, and their task is made much easier by unemployment and a falling cost of living. Industrial prices will decline at least to the extent of the reduction in wage and raw-material costs.

The process of price reduction is almost certain to have adverse effects on investment. Firms will tend to hold inventories to a minimum and fixed capital investments will be postponed in the expectation that capital-goods prices will be lower in the future. There is a tendency for a cumulative process to develop. The faster prices fall, the faster they are expected to fall. A stronger expectation of price cuts produces a further reduction in investment demand, which helps to speed up the price decline.

The change in the level of prices has mixed effects. A reduction in the price level increases the real value of money assets and increases the real burden of debt. If the two effects were symmetrical, the net effect on real expenditure would be favorable since money assets exceed debt in money terms to the extent of the sum of gold stock, treasury currency, and government debt. Professor Pigou has laid great stress on that point even though net money assets were, until World War II, a small fraction of total assets. But it is hard to believe that the oppressive effects of the increase in debt burden do not outweigh the favorable effects of the increase in real value of money assets even though the money assets exceed money debts. Households that hold assets stated in money terms for long-term purposes, e.g., retirement, will respond slowly and uncertainly to an increase in the value of their assets. They are not in the habit of adjusting their plans, such as they are, to changes in the price level, and, in any case, by the time they want to liquidate their assets the price level may have risen again. Debtors, on the other hand, must immediately adjust themselves to the increase in the burden of debt service or repayment obligations. The effects of an increase in debt burden are therefore likely to be considerably greater than the effect of a corresponding increase in the value of real assets. On the whole, then, it seems that the situation will become worse the greater the cumulative fall in the price level and the greater the rate of change of prices.

C. Security Values and Solvency of Financial Institutions

When incomes and prices fall, the values of marketable securities will also tend to fall. Because of fixed charges, profits will fall by a greater percentage than money income. As a result the values of marketable equities will also tend to fall by a greater percentage than income. That tendency may be reinforced by speculation on a further decline and by forced liquidation of securities. Once income has declined by a substantial amount, the deflation of security values will be extended to second-grade bonds as well as stocks. The leverage of overheads and the effects of speculation will reduce the value of the equity of the owners of unincorporated business and farms in the same way.

The propensity to consume tends to fall when the ratio of assets to income declines. In addition, the borrowing power of the owners of unincorporated businesses will be impaired by the decline in asset values. Those tendencies will, of course, reduce total expenditure and cause further reductions in asset values.

A decline in asset values may have an additional adverse influence on the movement of income through its effects on financial institutions. Financial institutions holding second-grade bonds or mortgages on property whose value declines may become insolvent.

When they cease to meet their obligations, assets which were regarded as liquid suddenly become frozen (and of problematical value). The amount of assets which become frozen may be much greater than the reduction in the value of the portfolio of the financial institution. Both consumption and investment will, of course, be reduced when financial institutions fail. In addition, "sound" financial institutions will become very cautious when other institutions fail.

The importance of those considerations will vary greatly with the initial situation and with the extent of the decline in income and prices. Security values may be relatively low at the start of the downswing because investors have been afraid of a depression or because of a lag in the adjustment of values to an increase in income. In that case, security values may not fall much and may even rise when income falls by a small amount. It may require a substantial reduction in income before equity values begin to fall very rapidly.

If, on the other hand, equity values at the top of this boom have been based on optimistic projections of future earnings or simply on the fact of rising security prices, a small reduction in income will bring on a very rapid reduction in the values of equities.

In the same way the vulnerability of financial institutions to reductions in income and prices varies with the character of their investment policy.

D. Inventory and Output Adjustments

Net investment in inventories is, on the average, a relatively small part of total investment. But in the short run, inventory investment is much more volatile than any other type of investment. Variations in inventory investment account for a very large proportion of the year-to-year variations in total investment.

In the long run, additions to inventories tend to be proportional to the average rate of change of output though the proportionality factor may change slowly with the trend in production and distribution techniques and competitive conditions. But in the short run, *desired* additions to inventories tend to change not only with the rate of change of output but with the relation of rate of change of output to rate of change of potential output (whether the limitation on output is labor supply or plant capacity). When output is high relative to capacity and is increasing rapidly, firms wish to increase the ratio of raw-material inventories to output to ensure themselves against slow deliveries. To the extent that they succeed in building up inventories they will have excessive inventories if the rate of increase of output declines. Desired inventory investment may therefore become negative even though the rate of change of output is still positive or zero. Similarly, when output declines, there may be a desire to reduce the ratio of inventories to sales because supplies can be quickly obtained (and to avoid obsolescence on style goods). Those tendencies will often be reinforced by the movements of prices and by changes in expectations.

Actual changes in inventories will deviate from planned or desired changes for several reasons. When capacity is strained, firms which order materials for the purpose of increasing inventories may not receive what they ordered. Firms generally may attempt to build up inventories by ordering or producing more than is required to balance current sales. When that happens, income is made to rise, sales exceed expected sales, and the actual addition to inventories is less than the planned addition. When firms are trying to reduce inventories they are partially frustrated by the reduction in demand which they cause. Attempts to increase (decrease) inventories do not ordinarily create (destroy) enough demand to balance the increase (decrease) in production which they cause, but it is important to note that attempts to change inventories may have a much greater influence on production than is indicated by reported changes in inventories. Moreover, when other factors cause aggregate demand to move in the same direction as inventories, attempts to adjust inventories may be completely frustrated for a time. The effort to change inventories will have its effect on production and income nonetheless. In 1930 and 1931 there was very little reduction in inventories. The actual reduction occurred in 1932 and 1933.

A matter closely connected with inventory adjustment is the direct influence of short-run demand expectations on production and thereby on demand. We usually suppose that production during any period will equal sales in the preceding period plus or minus planned inventory investment. We therefore suppose that, in addition to a relatively short lag between the receipt of income and the expenditure, there is a lag between the expenditure and the generation of income which is on the average much longer. The length of the "period" used in Keynesian sequence analyses is essentially the average length of time required for an increase in expenditure to lead to an increase in wage and other incomes. We usually equate income earned with the sum of expenditure by investors, consumers, and governments in the same time period.

As a definitional matter, that is correct if we count unplanned changes in inventory (positive or negative) as part of investment. But the investment used in sequence analyses is planned investment, not actual investment. Nonetheless, in most sequence analyses we still say that income earned equals consumer expenditure plus government expenditure plus (planned) investment. That procedure is justified in the following way. Assume that there is a very short lag between changes in income and changes in consumer expenditure (aside from longer lags explicitly introduced). Also assume that production in any period equals sales of the preceding period and that there is a lag between consumer expenditure and manufacturers' sales (there is actually a whole series of lags, but for simplicity we can speak of a single average lag). Then, neglecting the short lag between receipt of income and expenditure, we say that sales of consumers' goods at time t is determined by income at time t. Production of consumers' goods at time $t + 1$ equals sales of consumers' goods to consumers at time t plus planned inventory changes. Income generated by production of consumers' goods at $t + 1$ equals value of production of consumers' goods at time $t + 1$.

Aside from the simplification involved in averaging the lags, that is a correct procedure and leads to the same equations as those usually used. However, it rests on the assumption that firms always expect to sell tomorrow what they sold today and that the order lag is constant. As a rule of thumb, that is certainly a useful generalization, but deviations from it are sometimes important. It will sometimes happen that producers will expect sales to rise or fall from their current levels. They will then adjust production accordingly, not to change inventories but to avoid undesired changes in inventories. If a large number of firms expect the same kind of change in sales, their forecasts will be at least partially fulfilled. If households share the expectations of firms and raise consumption relative to income, each party will help to fulfill the other's forecast. When everyone expects income to increase rapidly and acts accordingly, a certain increase in income will be generated which will be added to whatever increase or decrease would have occurred without the change in expectations. Professor Burns has criticized Keynesian theorists for their neglect of considerations of that sort, and it seems to the writer that he is correct.

In a considerable range of variation in expectations the Keynesian view may be essentially correct because firms use their inventories to provide a buffer between changes in sales and changes in production. But when income begins to move upward or downward rapidly for other reasons, adjustments in production on the basis of projected changes in sales will reinforce the tendency for income to change rapidly.

The combined effects of cfforts to reduce inventories after income has fallen and to anticipate further reductions in sales can greatly accelerate a downward movement in income once it gets under way.

VIII. GROWING INDUSTRIES

When the aggregate income is growing, deviations of the rates of growth of particular industries from the average rate of growth make little difference, unless the deviations are correlated with the investment and borrowing characteristics of different industries. That may not be true when aggregate income begins to fall.

Suppose that aggregate income falls by a few per cent and that all firms find their sales reduced by the same percentage. As a result the returns from investment will fall generally, and so will gross profits. Gross investment will fall (inventory investment becoming negative), while dividends will be maintained for a time. The result may be that, on the average, gross investment plus dividends and withdrawals becomes less than gross profits, and business firms accumulate cash, though on a relatively small scale. That was what happened in 1930.

Now suppose, instead, that the share of aggregate demand going to some industries is growing rapidly. The demand for those industries will continue to grow in the face of a decline in aggregate demand. Output in some other industries will therefore decline faster than gross national product. It may then happen that the relatively declining industries are unable to accumulate cash, while the relatively growing industries continue to borrow. The difference between gross business investment and gross business saving will then be positive, whereas in the other case it was negative.

That is certainly not a necessary result and, indeed it is not a very likely one in general.

However, there is some reason to think that, during the nineteenth century, differential growth rates had an effect like the one described above. During the 1870s and the 1890s Midwestern agriculture continued to develop in the face of general depression. The expansion of agricultural output in the relatively new areas intensified the difficulties of Eastern farmers. Moreover, the interregional shift in income created investment opportunities in the trade and service and agricultural processing fields in the newly developed areas and cut the earnings of firms in those fields in older areas. A large part of the investment in the new areas was financed by borrowing rather than out of profits because new firms were involved. Investment in the older areas was reduced by the shift, but so was business saving. The net effect of the shift was therefore favorable to the maintenance of income.

During the 1890s the continued growth of capital-intensive electric utilities and street railways had a similar effect.

IX. SUMMARY AND CONCLUSIONS

The argument of this chapter can be summarized in the following way: (1) Normally the interaction of a capital-adjustment process with autonomous investment tends to cause real income to grow; (2) the upward movement can be interrupted by a number of different types of event; (3) these events either (*a*) prevent income from growing, so that capital stock rises relative to income or (*b*) cause a downward shift in the investment-demand functions of sufficient magnitude to cause income to fall; (4) in either of those cases, capital stock rises relative to income, which causes further reductions in investment and may lead to further reductions in income; (5) the resulting cumulative process is reinforced by adverse expectations, declining security prices, and some other factors; (6) the speed of the cumulative process is reduced by various lags in the system, by reductions in the ratio of business saving to income, by increase in population, and by a decline in government saving; (7) eventually the investment demand function will shift upward again because of the elimination of the original cause of the downward shift; (8) the upward shift in investment demand may produce a recovery, but whether it will or not depends on how much increase in the ratio of capital to income has taken place as well as on sensitivity of saving and investment to increases in the capital-income ratio. The increase in the capital-income ratio depends on the time interval between the start of the downswing and the upward shift in investment demand. It also depends on the speed with which income moves down. The speed of movement of income depends on the balance of the effects of the factors mentioned under 4, 5, and 6 above.

If the condition for a recovery is not met at the time when an upward shift in investment demand occurs, or if expectations have deteriorated so much that no shift does occur, income will have to fall until negative net investment is balanced by negative saving.

What has been said so far applies in a general way to all the different causes of downswing. But the different causes differ in the size of the initial shock they produce and in the length of time required before recovery can begin.

CHAPTER 12

Fluctuations Continued

In the last two chapters we have attempted to survey the factors which determine the rate of growth of aggregate demand and the factors which influence the response of the system to various types of adverse shocks. In Chapter 10 it was argued that the growth of income results from a rather complex interaction of a capital-adjustment process (or generalized multiplier-accelerator process) with exogenous factors, such as population growth and autonomous investment. It was argued that an interaction process of that type is theoretically capable of producing sustained growth of aggregate demand. But we also concluded that the system is not likely to be capable of sustained growth without the aid of population growth and some autonomous investment or some growing capital-intensive industries.

Depressions, it was said, may arise either because the system is incapable of sustained growth or because of the effects of adverse shocks. In Chapter 11 we assumed that in the absence of adverse shocks the system would be capable of sustained growth. We then considered the various factors which determine the amplitude and duration of depressions resulting from adverse shocks. The conclusions summarized at the end of the chapter were rather general and qualitative. In particular, they did not indicate how much depression results from a shock of a given type. Obviously we cannot reach any conclusions of that type without a numerically specified system. At best, what we have done is to give a guide or road map to the analysis of particular situations.

In this chapter we shall move a step toward concreteness by making some judgments on the adequacy of ceiling theories as explanations of major depressions. We shall offer some judgments as to reasons for the special virulence of the depression which followed 1929.

It will be argued that ceiling theories are not adequate explanations of the major depressions which have occurred in the United States. It will also be argued that the 1929 depression was much worse than its predecessors because of changes in the economic structure which reduced the stability of the system.

Those are, of course, empirical conclusions. Unfortunately they are supported only by qualitative evidence and by a good many judgments with which the reader is entitled to disagree. They must be taken, then, as hypotheses rather than as firm conclusions.

I. CEILING THEORIES

When income rises rapidly, supplies of labor, credit, or capital goods may not rise as rapidly as income. The rate of growth of income will fall when a ceiling is reached. If capital and income were growing at the same rate before, the rate of growth of capital will rise above the rate of growth of income. The ratio of capital to income will rise and cause a secondary reduction in investment. This may start a downward cumulative process.

In fact that does not seem very probable. In the nature of the case, limitations on the labor supply, the money supply, or the supply of capital goods will become important only when income has been rising very rapidly. The stabilizing influence of the consumption, dividend, replacement, and housing-construction lags will be very substantial at the moment when the growth of income is checked by ceilings. At the same time, prices will not fall rapidly if income merely stops growing or falls slowly. As we have already seen, expectations of income change will have only a moderate influence at first. If we rule out stock-market speculation, there is no reason for the position of financial institutions to change.

A fall in the rate of growth of income as a result of shortages will certainly cause a fall in the rate of inventory investment and is quite likely to make firms attempt to reduce inventories. But other types of investment will continue at high levels, and consumption may even increase. Attempts to adjust inventories will therefore be successful, and inventory decumulation will come to an end after a year or so.

The position of the system after ceilings have been operative for a year or so will be only a little different from its position when the ceilings first became significant. There will be, say, 4 per cent more fixed capital and a little less inventory. Income will not have fallen more than 2 or 3 per cent (and may not have fallen at all). Capital-income ratios will have risen only a few per cent. The ratio of fixed capital investment to income will fall, but by a relatively small amount, and the savings ratio will also fall. It is very likely then that income can begin to rise provided that the influence of ceilings has been removed.

When we consider the nature of the ceilings which hold down the growth of income, it seems obvious that their influence will ordinarily last for only a short time.

A shortage of capital goods is, in the nature of the case, a temporary phenomenon. When investment demand increases very rapidly, capital-goods producers may be slow to expand plant. That may be so because they doubt the persistence of the boom, because they lack funds for a very rapid expansion, or simply because it takes time to build plant and to recruit labor. Ordinarily, however, producers of capital goods will be willing to build new plant when they reach a rate of output which is considerably less than 100 per cent of practical capacity. They may not respond quickly enough to prevent bottlenecks from occurring, and the growth of income may be checked by shortages of capital goods. But if the argument given above is correct, capital-goods producers will still experience very high utilization rates for a couple of years after investment has stopped growing. Usually they will expand capacity soon enough to permit income to start to rise again when the bottlenecks are removed.

A labor shortage which does not produce a wage-price spiral can be corrected in the same way as a shortage of capital goods. Suppose that income tends to grow more rapidly than the labor supply and its productivity. There may be some unemployment initially, but eventually labor shortages will develop. Wages and prices are likely to begin to rise before it becomes very difficult to recruit labor. The effects of changes in wages and price changes of that sort are discussed in Chapter 9. For the moment let us consider only the real effects of labor shortages.

When it becomes difficult to recruit labor, investment may be unaffected for a time, but the rate of increase of real income will have to fall to the rate permitted by the expansion of the labor force. If capital and real income were growing at the same rate before the shortages, capital will now grow faster than real income. The resulting rise in capital-income ratios will depress investment. The rate of growth of demand will fall, and an inventory cycle will temporarily reduce the growth rate to a figure below that permitted by the growth of the labor force. The labor shortage will then be ended. If at that point the rate of growth of capital stock has fallen below the rate of growth of income, there will be a recovery during which the rate of growth of income may increase until labor shortages again develop.

Of course, investment may be influenced directly by the labor shortage. Recruiting difficulties may make extensive investment less profitable. In that case the labor shortage will have effects which are very similar to those produced by a shortage of capital goods.

A shortage of credit will produce effects which are very similar to those produced by shortages of capital goods, provided that the money supply is rising secularly.

Ceilings on resources may help to explain short minor cycles, but it does not seem very likely that resource limitations *in themselves* will cause major depressions. Nonetheless, resource limitations may play an important role in bringing on major depressions. They may do so in a number of ways: (1) A reduction in the rate of growth of income may be the proximate cause of the collapse of a speculative boom. (2) A labor shortage may cause a wage-price spiral and a general speculative boom, and a capital-goods shortage may bring on a boom in the capital-goods industries. In 1873 rising prices for coal and iron contributed to the collapse of the American railway boom and also caused a boom (which later collapsed) in the development of British coal mines. (3) Shortages of labor or of funds may reduce the flexibility of the economy so that expansion in one sector cannot offset contraction in another.

Ceilings on resources are not in themselves likely to produce major depressions because they do not produce large reductions in the rate of investment and because a temporary reduction in the rate of growth of income is sufficient to remove their influence.

A. The 1953–1954 Recession

The argument with respect to ceilings depends in part on judgment as to the strength of the stabilizing tendencies in the economic system. The judgment made here is supported to some extent by the empirical evidence.

The recession of 1953–1954 was probably not caused by limitations on the labor supply or any other resource, but it does give us some idea of what happens when income stops rising. Since the essence of the ceiling theory is the notion that, when income stops rising, it must fall, we can use any recession to test the stability of the system.

From the second quarter of 1953 to the first quarter of 1954, gross national product (in current prices) fell from $367.4 billion to $357.6 billion—a decline of 2.4 per cent.[1] Gross national product then remained roughly constant for two quarters and then began to rise.

The decrease in GNP took the form of a reduction in government expenditures and a shift from positive to negative inventory investment. The other expenditure components showed great stability. Residential construction fell by 3 per cent from the second quarter of 1953 to the first quarter of 1954. It then began to rise, and by the third quarter of 1954 residential construction had increased 15 per cent over the second quarter of 1953. The increase was no doubt due partly to the easy-money policy of the government (and to changes in mortgage-credit regulations). However, the important point here is the stability of residential construction in the face of a decline in income and a continued rise in the housing stock.

Other private construction remained virtually constant during the last half of 1953 and rose somewhat during 1954.

[1] Data in this section are drawn from the *Economic Report of the President*, January, 1957.

Investment in producers' durables fell off by 12 per cent from the second-quarter of 1953 to the first quarter of 1955. Meanwhile, however, corporate profits (after inventory-valuation adjustment) fell from $39 billion to $29.8 billion, while dividends rose from $9.2 to $9.7 billion. (Corporate saving did not reflect this change because of changes in tax legislation.)

Finally, personal-consumption expenditures, after falling by less than 1 per cent in the last half of 1953, rose by about 1 per cent in the first half of 1954.

There is no doubt that the depression would have been worse without government action in the field of taxation and credit policy. Nonetheless, the data given above do indicate that the stabilizing tendencies discussed earlier really operate. Investment in construction was very stable, personal saving fell slightly, and the reduction in investment in durables was more than offset by the decline in corporate profits and the rise in dividends.

Once the adjustment in inventories was completed, a strong recovery began. It should also be noted that the reduction in government expenditures which took place in 1953–1954 was a shock of greater magnitude than any likely to result from the appearance of resource shortages.

The question remains whether the economic system has enough stability to weather the effects of resource shortages. At least it can be said that some stabilizing mechanisms are at work. In the nature of the case, shortages of resources are not likely to cause a very large or sudden drop in investment (unless they cause a wage-price spiral and a major inflation). The supporters of ceiling theories ought to produce at least one example in which resource limitations per se caused a major depression.

B. The Causes of Major Depressions

We may reject the ceiling theories as causes of major depressions, but depressions must have had some cause. A number of possibilities have been suggested here. No one of them is likely to be the exclusive cause of any particular depression, but it may be worthwhile to mention some cases in which the causes discussed above have played a role.

In 1873 the rate of railway construction in the United States declined very sharply. It is not necessary to maintain that the sudden decline in railroad construction was the sole cause of the depression of 1873, but it was certainly a remarkable coincidence if the railway depression had nothing to do with the general depression.

The somewhat milder contraction of 1883–1885 was also associated with a decline in the rate of railway construction.

The situation in the depression of 1893 is more difficult to assess. The most obvious factor in the situation was the financial crisis. That crisis apparently arose from a general apprehension that the United States would be forced off the gold standard by its silver-purchase policy. The gold crisis was over by the beginning of 1895, but the Bryan campaign and the war scare of 1896 caused a renewal of the crisis. One never knows how much weight should be given to purely monetary factors of that sort. It can always be argued that a depression was bound to occur anyway, so that the credit crisis was the proximate but not the fundamental cause of the depression. Perhaps so, but the beginning and end of the crisis coincided with the beginning and end of the credit crisis. It seems hard to believe that the credit crisis did not play an important role in bringing on the depression. Of course, other factors must have contributed.

The most simpleapproach to the depression of 1907 is to regard it as an aggravated case of the minor monetary crises which chronically afflicted the American economy under the National Banking System. The aggravation consisted in the fact that ten years of rising income and prices had created a stock-market boom and a boom in new issues of industrial securities. The depression, though severe, lasted a very short time. It is therefore hard to find another explanation. No structural maladjustment could be corrected in so short a time. Autonomous investment could hardly move downward so violently and recover so quickly.

The depression of 1920–1921 seems to be a clear-cut case of the collapse of a speculative boom based on a wage-price spiral. A number of causal factors associated with the depression of 1929 have already been cited.

These *ad hoc* explanations are bound to seem unsatisfactory to anyone who is in search of a systematic cycle theory. The only cyclical element in the theory outlined here is the fact that it takes several years of prosperity before any of the causes of depression mentioned here can become operative.

To make a really sharp test of the theory outlined here, one would have to make a very detailed study of the course of each historical cycle, studying not only the aggregate data but the saving and investment behavior of each sector. It can only be said here that the explanation of business-cycle history offered here is in line with the more obvious facts of economic history.

One may cite the stock-market boom and the collapse of speculative building of apartments and office buildings as the proximate causes of the depression which began in 1929. Those events were not so very different from the events which brought on the earlier major depressions, but the depression which followed was certainly different.

II. THE 1929 DEPRESSION

The events of the years following 1929 showed how unstable our economic system can become. From 1929 to 1933 (real) gross private domestic investment fell by 90 per cent, while real gross national product fell by 30 per cent. It was about eight years before GNP reached the 1929 level.[2] By contrast the great depressions which began in 1873 and 1893 were much milder. Investment (as measured by Frickey's index of durable-goods productions) fell by 33 per cent from 1872 to 1876 and by 34 per cent from 1892 to 1894. Frickey's index of manufacturing production fell by only 10 per cent in the depression of 1873 and by 15 per cent in the 1890s.[3]

The other "major depressions" which began in 1883, 1907, and 1920 were both shorter and milder. Why was it that income and investment could decline as much as they did in the 1870s and 1890s without sending the economy into a tailspin? Why was the depres-

[2] *Ibid.*

[3] Edwin Frickey, *Production in the United States, 1860–1914*, Harvard University Press, Cambridge, Mass., 1947, pp. 54, 64.

sion of the 1930s so much more severe? Any adequate business-cycle theory ought to give an answer to those questions.

It has been argued here that certain types of change (or difference) in the structure of the economic system contribute to its stability, while changes in the opposite direction contribute to instability. We shall now argue that the historical changes in the structure of the American economy which occurred during the first quarter of the twentieth century tended to reduce the stability of the system. In addition, some of the consequences of World War I caused a further reduction in the stability of the system. We shall try to show in a rather rough-and-ready way that those historical changes account for the difference between the behavior of the system after 1929 and its behavior in the 1870s and 1890s.

The stability of the American economy was lower in the years after World War I than it had been previously for a number of reasons: (1) The importance of rapidly growing sectors of the economy was smaller after World War I than in earlier years. (2) The backlog of housing demand which gave some support to investment during the depressions of the nineteenth century was very low in 1929. That was due to relatively easy mortgage credit and to speculation in the construction of apartments in the 1920s. (3) The rate of population growth was lower after World War I than in earlier decades. (4) The stock-market crash in 1929 had a more direct impact on consumption than the previous financial panics. In the next section each of these factors will be discussed in turn.

Before going on, it should be remarked that the factors just given are not intended to constitute an exhaustive list of the factors involved in bringing on the 1929 depression. It is a list of the more obvious differences between 1929 and earlier crises. Of course, many additional factors played a role in both cases. For example, investment in the automobile industry and many corollary industries, i.e., petroleum extraction and refining, garages, and service stations, involved building ahead of demand. The rate of growth of the automobile industry was probably due to taper off by 1929. In some respects, then, investment in the automobile and related industries played a role analogous to the role of the railroad industry in the late nineteenth century. However, its relative importance was probably not so great.

A. Growing Industries

In the period between the Civil War and World War I a number of sectors of the American economy went through periods of rapid development, in each of which the sector in question tended to grow relatively to the rest of the economy. At any one moment during the period after the Civil War a large agricultural area had reached the stage of producing large amounts of output for export and created a basis for investment in strongly market-oriented service trades—retailing, wholesaling, local transport, and agricultural processing. At the same time somewhat older areas had acquired a large enough population in agriculture and the agricultural service trades to attract investment in market-oriented manufacturing or in those branches of manufacturing which are partly market-oriented and partly oriented to the natural resources available in the area in question.

The development of new areas contributed to the growth of demand in two ways. (1) It led to increased exports of agricultural products. (2) There was a considerable lag between the time when investment expenditures were made and the time when the resulting output appeared. The frontier areas therefore provide an example of the effects of investment made in anticipation of sales which were discussed in Chapter 10.

The development of manufacturing in the Midwest had a similar effect. Similar developments took place in the railroad field. The great transcontinental lines were built ahead of demand, and so were many less spectacular ventures.

By the late 1880s, investment in railroads had begun to decline, but by that time investment in street railways, electric utilities, and telephones had begun. The facilities for those industries had to be built ahead of demand, and they were also very capital-intensive.

There cannot be much doubt that those developments contributed substantially to the growth of demand during the period between the Civil War and World War I.

The development of new areas, new types of public utilities, and new manufacturing industries also influenced the stability of the system. The speculative character of railroad investment was, of course, a destabilizing influence, but other types of autonomous investment had a stabilizing influence. The different types of autonomous investment did not all move together. Railroad investment fell off sharply after 1873, but the development of Midwestern agriculture continued at a rapid pace. Grain output in the Midwestern areas increased rapidly throughout the depression of the 1870s. In spite of the decline in grain prices, the aggregate income of the Midwestern area must have increased during the depression. That increase in income must have generated a substantial amount of investment.

Of course, the gains of the Midwest were partly offset by losses elsewhere. Farmers and their suppliers in other areas lost income when grain prices fell. Naturally they invested less than they would have invested if there had been no competition from the new areas.

But the gains and losses did not balance. A large part of the positive investment was probably financed either from borrowing or from previously accumulated liquid assets. On the other hand, the reduction in investment in the older areas was probably balanced largely by a reduction in saving.

To put it another way, the increase in Midwestern agricultural output at a time when aggregate money expenditure was not rising reduced the aggregate real income of farmers. The real income of nonfarm consumers was therefore increased (by comparison with what it would have been). Farmers appear to have a much higher short-run marginal propensity to consume than urban consumers. Therefore aggregate savings tended to fall. At the same time investment was maintained by the development of the Midwest.

Regional shifts and the adoption of new techniques in manufacturing probably had analogous effects. During the 1890s investment was supported by the shift of demand toward the very capital-intensive public utilities.

It is fairly obvious that the forces supporting autonomous investment in the 1920s were much weaker than those at work in the late nineteenth century and in the years before World War I. Frontier development had been completed. Investment in the public-utility and transport group of industries was much smaller (relative to GNP) after World War I than in the earlier period.

The 1920s were, of course, marked by the development of the automobile and related industries, but it was not necessary to build so far ahead of demand in those industries. Internal finance played a much more important role in the development of the automobile industry than in the public-utility industries.

B. Population Growth

The stability of the economy of the late 1920s was also lowered by comparison with the nineteenth-century economy by the change in the rate of population growth. It was shown above that population growth contributes to stability in two ways. (1) Population growth increases the rate at which the personal-savings ratio declines when income falls. (2) Population growth reduces the rate at which the demand for housing falls as aggregate income falls. The rate of population growth was 32 per cent per decade from 1870 to 1880, 21 per cent from 1890 to 1900, 16 per cent from 1920 to 1930, and only 8 per cent from 1930 to 1940. Obviously, then, the population factor contributed much more to both growth and stability in the nineteenth century than after World War I.

C. Residential and Commercial Construction

The rapid rate of population growth and the rapid development of new industries and new areas in the nineteenth century tended to produce a high rate of growth of income. Those forces had declined in relative importance by the end of World War I. Other things equal, we might have expected to find that the rate of growth of income during the 1920s was lower than in prewar decades. That was not the case (to any marked extent).

One factor which helped to maintain the rate of growth of income during the 1920s was the very high rate of residential construction.

The rate of growth of income was supported during the 1920s by four aspects of the position of the housing industry: (1) the shortage of housing at the beginning of the decade, brought about by the cessation of building during World War I; (2) the transitional effect of the change to easier credit conditions; (3) the permanent effect of easy mortgage-credit conditions; (4) the speculative boom in apartment and commercial buildings. All of those factors except the third were temporary in character. Not only were they temporary but, when their favorable effects were exhausted, they had a temporary negative impact on investment and the rate of growth of income. The third factor, though not cyclical in character, had an adverse effect on the stability of the system when the depression came.

The positive effect of the initial shortage of housing on investment and the growth of income is obvious. Its temporary nature is equally obvious—the backlog of housing demand was bound to be worked off eventually. In much the same way the easing of credit conditions created a "one-shot" increment in the demand for housing which stimulated building temporarily. At the time of the change in credit conditions there existed a certain group of families whose incomes were large relative to their assets and who would normally have bought homes at a later date. The reduction in down payments enabled them to buy homes earlier than they otherwise would have done. At the time when they would normally be in the market for housing, their demand was already satisfied. Their places, however, were taken by others. Demand was therefore temporarily increased during the period in which the brought-forward demand was added to the normal demand. Once that temporary increase in demand was satisfied, demand returned to normal for given income and population.

Neither the initial housing shortage nor the temporary burst of demand due to the change in credit conditions had a reaction in the form of a reduction of demand to a level below the norm for given income and population. They did produce an adverse reaction through their effects on the rate of construction. It is shown in Chapter 7 that a housing shortage will produce a cyclical movement in construction. In order that a housing shortage be met, the rate of increase of housing supply must become higher than the rate of increase of demand generated by normal increase in income and population. The building industry is bound to overshoot the mark and create a surplus after it has met the shortage.

The speculative boom in apartment and office building was obviously temporary in character and, unlike the other temporary factors in the situation, had a direct negative effect on construction when it was over. Because of the size of the units, the long period of construction, and the interest of promoters in prolonging the boom, it is possible to generate a much greater excess supply of apartments and office space than of single-family houses. The slump which follows the boom is therefore much more severe in the large-scale-building field. Moreover, apartment- and office-building activity is heavily concentrated in a few large cities, so that the local booms and slumps have relatively little chance of canceling each other out.

Apartments and offices were overbuilt even before the 1929 crash, so that investment in this field would have declined sharply even had there been no reduction in income in 1929. When changes in expectations or financial conditions make firms less willing to invest than before, the rate of change of income tends to fall or to become negative, and capital tends to accumulate faster than income, which has further adverse effects on investment. The speed of that cumulative process depends inversely on the extent to which demand for capital is generated by factors other than the growth of income and the extent of the backlog of capital demand already in existence. From this point of view, overbuilding in the apartment- and office-building field clearly tended to increase the speed of the downward cumulative process. In the same way the relatively smaller amount of overbuilding in the single-home field was destabilizing. Finally, the easy-credit situation was destabilizing because it reduced the reserve of demand for single houses in existence at the end of the boom.

D. The Effect of the Stock-market Boom

The great stock-market boom of the 1920s had a number of effects on the real system which are difficult to disentangle. There was a redistribution of the flow of funds through financial markets as the result of an interaction of the stock-market boom with other factors in the situation. The high and rising prices of equities encouraged corporations to sell more equity issues and make fewer bond issues than they would normally have done. At the same time, changes in corporate financial policy (partly as a result of the experience of 1920–1921) led to a tendency for corporations to finance a larger part of working-capital needs by raising long-term money with a corresponding reduction in the demand for bank loans. A large part of the increase in demand for equity money was lent by banks through the channel of brokers' loans. Banks lent to brokers, who lent to margin buyers of stocks. Some of the sellers of those stocks may have hoarded the proceeds, but most of them went back into the market, and a large part of the funds provided by banks, after being kicked around the market for a time, went into new issues. At the same time some of the funds which would normally have gone into industrial-bond issues went into the purchase of real-estate bonds. Thus the stock-market boom played a role in bringing about real-estate speculation.

To the extent that the boom enabled corporations to raise funds from equities instead of from bond issues or bank loans, the stability of the system was increased. But the improvement in the financial position of corporations was offset by the worsening of the position of households. Households which indirectly financed new issues by borrowing to buy stocks on margin were left in debt to the banks or gave up bank deposits to meet margin calls. As the bank loans were repaid a corresponding amount of deposits was extinguished.

The balance of the effects on corporations and households was definitely unfavorable. Large corporations whose position was relatively strong anyway were able to accumulate large amounts of cash during the downswing instead of having to pay back bank debt. They also had a smaller volume of long-term debt outstanding than they would have had without the market boom. The favorable effect of those factors was probably small by comparison with the unfavorable effect of the loss of liquid assets and the burden of short-term debt on households. Moreover, some of the losses of households were matched not by gains of industrial corporations but by worthless real-estate securities. It is impossible to estimate the magnitude of the effects of the market boom and crash on consumption and investment, but it must have been substantial.

The stability of the system was also weakened by the positive effects of the boom. Easy credit to both firms and households certainly increased the rate of growth during the boom, but it also reduced the backlog of investment demand for businesses and the backlog of demand for durable goods for households.

Some notion of the joint effect of that consideration with the loss of assets to households can be obtained from the size of the decline in consumer purchases which took place immediately after the market crash.[4]

E. Other Factors

A final support to the rate of growth of income during the 1920s was foreign lending. It is obvious that foreign lending contributed to the income stream. When it came to an end, little was taken out of the income stream, since little was repaid, so that it had no direct adverse effect on the system. But when the foreign securities went into default, another blow was struck at the asset position of the households and banks which held the bonds.

To complete the sorry tale, the banking system was demoralized when its holdings of dubious mortgages and real-estate bonds became worthless or frozen. This demoralization had an adverse effect on expectations generally and, in addition, tied up at various times from $2 to $4 billion of deposits.

No one will ever figure out the relative importance of all the factors at work in the years immediately after 1929 or know which was the straw that broke the camel's back. The situation in 1929 contained almost every special destabilizing element that had ever

[4] See Harold Barger, *Outlay and Income in the United States*, National Bureau of Economic Research, New York, 1942.

In 1929, consumers' outlay fell by $900 million (or almost 5 per cent) from the third to the fourth quarter. Income was falling during the fourth quarter, so that only part of the reduction in consumption was autonomous. But such a large and rapid change in consumption cannot be attributed to the reduction in income.

appeared before. In addition, the relatively long-term factors, such as the development of new areas and industries, contributed much less to the stability of the system than they had done before.

F. Was Sustained Growth Possible?

Try to imagine the 1920s without the stock-market boom and foreign lending, without a speculative boom in apartment and office building, without a postwar housing shortage, and, to make things even, without a debt-ridden agriculture. Such an economy would have had a substantially lower rate of growth than the one observed during the 1920s, and its growth would have been lower than the growth observed in earlier decades. How much lower the rate of growth would have been we cannot know. Investment in housing would have been lower without the temporary stimuli provided by the initial housing shortage, speculation, and easy credit. It would have been lower still because income would have grown more slowly.

At best, income would have grown throughout the period at a rather low rate. But it is quite possible to believe that capital would have accumulated faster than income, with the rate of growth of income slowly declining until it became negative and the economy entered a deep depression.

It would take more information than we have now or are ever likely to have to find out what might have been. That is an unfortunate state of affairs. The answer to the question posed above must be a major factor in deciding whether, to put it simply, a boom is a good thing or a bad thing.

If the economy of the 1920s was fundamentally able to maintain a positive rate of growth without artificial stimulation, we can clearly say that the depression was essentially due to the "excesses" of security and real-estate speculation and to easy mortgage credit and foreign lending. Of course, those excesses would not have caused such a severe depression if the long-term factors governing stability had not also changed adversely.

But if the system was bound to work itself into a depression anyway, it is not clear that the excesses were so bad. They certainly made life merrier, and it is not at all clear by how much they shortened it.

III. SUMMARY

It was argued above that (1) ceiling theories do not give a very plausible explanation of major depressions; (2) such depressions can be explained by a combination of factors (different in each case), including the aftereffects of various types of speculative booms and fluctuations in autonomous investment; and (3) the American economy had a greater degree of stability in the nineteenth century than it had in the 1920s. The difference was due in part to a variety of changes in the structure of the economy and in part to a combination of temporary factors. The simultaneous occurrence of a number of adverse shocks was probably due to World War I. The housing boom and the stock-market boom, the boom in foreign lending, and the depression in agriculture would not have occurred together in such a short period had it not been for the World War.

The whole argument of this chapter is somewhat speculative in character. Its main empirical support comes from the fact that it does explain some facts which are not too easily explained by other approaches.

CHAPTER 13

Wages and Labor Supply

Some years ago Mr. Harrod[1] suggested that capitalist economies must have either cyclical unemployment or chronic unemployment. If, to employ his terminology, the "natural rate of growth" exceeds the "warranted rate of growth," there will be chronic unemployment. On the other hand, when the warranted rate of growth exceeds the natural rate, depressions brought on by labor shortages are inevitable. Mr. Harrod pitched his argument a little strongly by using an income-generation model of a very unrealistic type. Nonetheless, he raised a very real problem. If we give up our faith in flexible wages and interest rates and a high elasticity of substitution between capital and labor, it is certainly not obvious that the growth of demand for labor can be brought into line with the growth of the labor supply.

Let us suppose for the moment that the equilibrium rate of growth, in our terminology, exceeds the rate of growth of real income permitted by the growth of labor supply and its productivity. Then, according to the argument of the last chapter, there will be a series of relatively mild short cycles. In that way the average rate of growth can be brought into conformity with the natural rate. In each cycle, business saving, and possibly personal saving as well, fall below their equilibrium values. In that way the average savings ratio is brought into line with the rate of growth of income and the ratio of capital to income. The ratio of capital to income oscillates but does not necessarily have a trend. That is essentially the same mechanism used by Harrod to produce consistency among rates of growth, savings ratios, and capital-output ratios. According to the argument made in the last chapter, labor shortages may produce a

[1] R. Harrod, *Toward a Dynamic Economics*, Macmillan & Co., Ltd., London, 1952.

series of short, mild depressions every three years or so instead of one big depression every ten years or so.

However, that is not the only way in which the rates of growth of demand for and supply of labor can be equated. Wage rates have some tendency to rise before labor shortages have had a very serious effect on the rate of growth of output or on investment. If the money supply does not grow as rapidly as money income, investment will be held down by limitations on the supply of credit. Mild cycles similar to those resulting from labor shortages may result.

Alternatively labor shortages and easy money may generate a wage-price spiral and speculative boom which will ultimately break and cause a depression of major proportions. We shall argue, however, that the wage-making process is such that rapid inflation need not result from a labor shortage even when money is easy.

Finally, it is possible to have a relatively low unemployment percentage for a long time without any labor shortages. It was argued in Chapter 10 that the equilibrium growth rate of economic systems like ours is not likely to be very high. In Chapter 11 we were mostly concerned with stability in depressions, but nearly everything said there applies in reverse. It is difficult to produce very high rates of growth even temporarily. We cannot suppose that, whenever the actual rate of growth of demand exceeds the equilibrium rate, income tends to rise progressively faster. It is not at all obvious that the growth of demand is going to outrun the growth of the labor supply in every boom. It may, but there is no necessity for it to do so. If that were the whole story, we should still have to have either unemployment or labor shortages. For it is not very likely that the rate of growth of income will ever exactly equal any predetermined figure. But we shall argue below that the rate of growth of effective labor supply is determined partly by the ease with which jobs can be obtained, i.e., by the amount of employment and the number of job vacancies. Demand and supply for labor can be brought into rough equality by adjustment on the supply side as well as on the demand side.

In this chapter we shall discuss the processes determining the rate of change of wages, the relation between labor supply and employment, and some aspects of labor mobility which influence the size of the effective labor force.

In the first section of this chapter we shall discuss the process of wage determination. The purpose of this discussion is twofold: (1) It provides a basis for a discussion of the relations between the rates of growth of real income and money income and the growth of money supply, which is the subject of Chapter 14. (2) It is shown that the wage-making process does not tend to equate wages for similar labor in different sectors into economy. That makes it possible to consider the hypothesis that labor shortages can be reduced or eliminated by shifting labor from low-wage–low-productivity sectors to high-wage–high-productivity sectors.

In the second part of this chapter we consider a number of different processes by which the effective labor supply can be adjusted to the demand.

I. WAGE DETERMINATION

The theory of wages has been in a somewhat chaotic state for a number of years. The complexities of trade-union negotiations create a host of problems which do not exist in competitive markets. But even aside from the problems raised by trade unions, the classical theory of wages has taken a bad beating at the hands of labor economists. However one may interpret the marginal-revenue-product theory of demand for labor, one cannot account for the observed dispersion of wages paid to essentially similar workers in markets in which trade unions are not the dominant factor. Even in the absence of trade unions, labor markets do not operate like the markets of classical theory.[2]

It will be a long time before we have a theory of wages which takes account of the complexities of trade-union policy, the determinants of labor mobility, the wage policy of employers faced with

[2] Richard A. Lester, "Wage Diversity and Its Theoretical Implications," *Review of Economics and Statistics*, vol. 28, 1946.

Joseph Shister, *Economics of the Labor Market*, J. B. Lippincott Company, Philadelphia, 1949, p. 424 ff.

uncertainty and oligopolistic competitors, and the limitations on intraplant wage structures imposed by considerations of morale. Meanwhile, however, it may be possible to single out a few aspects of labor-market processes which have a direct bearing on business-cycle problems. In this section we shall attempt to show in a rough-and-ready way why the rate of increase of wages tends to vary with the level of unemployment, why wage differences can persist, and why there is a tendency for labor to move from low-wage industries to high-wage industries when there is excess demand for labor in the aggregate.

It will be useful to begin by considering the nature of the forces bearing on the determination of wages and to look at them not from an equilibrium point of view but from a dynamic point of view. That is, we want to look at forces which tend to cause wages to change at a given moment or to prevent them from changing. We have to deal not with a single labor market but with a set of interrelated markets defined in terms of industries, locations, and types of job. There is a market for labor for a given type of job, in a given industry, in a given area. Some of the pressures on wages operate on all the plants in an industry regardless of location or type of labor. Some of them operate on all the plants in a given area regardless of type of labor or industry. Others work on all the labor of a given type in a given area regardless of industry.

The forces influencing wages fall into three groups: (1) those which tend to move a particular wage rate upward or downward independently of its relation to other rates; (2) those which resist the upward or downward movement of a wage rate independently of its relation to other rates; (3) a set of forces which might be called "binding forces," i.e., forces which tend toward the maintenance of a certain structure of wages, i.e., to certain relations between the wages paid for different jobs in the same plant, or between the rates paid by different firms in the same area for the same type of job, or between the rates paid by different firms in the same industry.[3]

[3] J. Dunlop, *Wage Determination under Trade Unions*, Augustus M. Kelley, Inc., New York, 1950 pp. 123–130.

A. Wage Pressures

The forces which encourage or resist wage changes (aside from the binding forces) may be roughly grouped into three categories: (1) supply and demand, (2) profits, and (3) standard-of-living pressures. As we shall see, these forces interact with one another in a complicated way. Excess demand for labor tends to push wages up, but excess supply of labor does not tend to push wages down. It does, however, weaken the resistance to other downward pressures on wages. Conversely, a high profit rate does not in itself tend to raise wages, but it weakens resistance to other upward pressures. On the other hand, a low profit rate does exert a positive downward pressure on wages. A reduction in living standards of workers resulting from a rise in prices relative to wages produces upward pressure on wages, while an increase in real wages does not lead to downward pressure but either weakens other upward pressures or weakens resistance to other downward pressures. Each of the forces mentioned operates through a number of different channels, some of which affect only one set of markets, while others operate over the whole range of markets.

The binding forces include: (1) Competition between firms, which limits the dispersion of wage levels among firms in the same industry. (2) Various ideas of a fair relationship between rates on different jobs in the same plant, which limit the range of variation of intraplant wage structures. (3) The establishment through trade unions or otherwise of "going rates" for types of labor employed by a variety of industries in the same local labor market. Together (2) and (3) link the wage levels of all the plants in a local labor market even though most of the jobs are not comparable. (4) Trade-union demands for uniform rates or systematically related rates over an industry operating in a number of areas link wage structures in different local labor markets. (Since more than one industry is subject to industry-wide bargaining, the resulting influences can be very complicated.)

Since we are interested in the relation of aggregate demand for labor to movements in the average level of wages, we shall confine our attention for the most part to the pressures on wages and neglect the influence of binding forces, although in principle both sets of considerations ought to be taken into account.

B. Supply-and-demand Pressure

A shortage of labor in the sense that unfilled vacancies are high relative to unemployment tends to lead to increases in money wages in two different ways. On the one hand, scarcity of labor (either in general or in particular occupations) may lead employers to initiate (or accede to union demands for) wage increases in the hope of improving their ability to recruit labor. On the other hand, tight labor markets lead to increased turnover rates, relaxation of discipline, and a greater intransigence on the part of labor in disputes over job content, etc.

It might be said that there is always a good deal of labor dissatisfaction which tends to reduce efficiency—which is held in check by the difficulty of getting a new job when unemployment is substantial. Job dissatisfactions manifest themselves more readily when jobs are easy to get and an employer may find it possible and profitable to offset increasing dissatisfaction by increasing wage rates (or fringe benefits). Though these two forces operate jointly, it will be useful to consider them one at a time.

1. *Wages and Labor Markets*

It used to be said that a single firm employing only a small percentage of the labor of a given type in a local labor market could be regarded as a perfect competitor in the labor market. It now seems clear that under full-employment conditions almost any employer has to be regarded as to some extent a monopsonist, i.e., faced with a rising supply curve for labor. Moreover, the supply curve's position depends very largely on the amount of labor he is now employing and the wages he is now paying.

This state of affairs arises from the fact that most workers are not in any real sense in the labor market. Most employed workers are not constantly on the lookout for jobs offering a better combination (from their own viewpoint) of wages and working conditions than the job they now have. Their knowledge of alternative combinations of wages and working conditions is not very accurate, though its accuracy varies with the character of the local labor market.[4] (Workers in a diversified, poorly unionized market tend to be very badly informed. Those in a highly unionized one-industry town may be very well informed. Even in the latter case, information on important aspects of working conditions, such as relations with foremen, may be difficult to obtain except by actual trial.) Even when workers do have all the relevant information, most of them are not likely to act on it.

When a worker has worked for a firm a certain length of time, he acquires certain advantages in staying where he is. His seniority gives him protection against layoffs which he would lose if he took another job. If he is a semiskilled worker, he may have to start on a lower rung of the intraplant job ladder than the one he now has if he changes jobs (or by losing seniority he may increase the length of time before he can move up). By changing he may stand to lose in terms of wages or working conditions even if the other plant pays a higher scale for given seniority. He may also lose pension rights. Even when those considerations are not relevant, workers do not respond very readily to every opportunity to improve their economic position. Workers are not seeking to maximize net utility (unless we stretch the sense of that expression a great deal).[5,6] Employed workers are likely to stay where they are so long as the job they now have provides an income which is above a certain minimum level and which does not involve more than a certain maximum of various types of dissatisfaction. There is, of course, a certain willingness to exchange dissatisfaction for income, but there

[4] Lloyd Reynolds, *The Structure of Labor Markets*, Harper & Brothers, New York, 1951, p. 210 ff.

[5] Reynolds, *op. cit.*, chaps. 2–5.

Richard A. Lester, *Adjustments to Labor Shortages*, Industrial Relations Section, Department of Economics and Sociology, Princeton University, Princeton, N.J., 1955, p. 68.

[6] Richard A. Lester, *Hiring Practices and Labor Competition*, Industrial Relations Section, Department of Economics and Sociology, Princeton University, Princeton, N.J., 1954, pp. 73–83.

are limits to the exchange. We might suppose that all types of jobs for which a worker qualifies can be divided into a satisfactory class and an unsatisfactory class.

If a worker is employed in a job in the acceptable range, he tends to stay in it even if jobs on higher indifference curves are available. If he is employed on a job in the unacceptable range, he tends to leave it (*a*) when he has reason to think that better jobs are available or (*b*) when he has enough resources to remain out of work long enough to find a better job before he runs out of resources and has to take a job in which his last state is worse than his first.

Workers who are not employed and have resources to permit them to stay out of work will refuse jobs within the unacceptable range and may, to a certain extent, turn down even acceptable jobs in the hope of finding better ones, but the length of time during which they will shop around is limited.

The standards of acceptability are not, of course, fixed. The standard of a satisfactory income is in part a function of past experience, and there will always be some tendency for workers to get used to what they have. Workers who were at one time forced to take a job which substantially reduced their living standards will gradually get used to their new positions and will not be so ready to leave as they were at first. In the same way one may get used to various types of job dissatisfaction. On the other side, workers who were fortunate enough to find jobs which raised their standards will try to hold out for an equally high income if they ever become unemployed again. Individual workers' standards will, of course, also vary with their family responsibilities, while willingness to leave a job will vary with age and family responsibility.

Workers' standards of income satisfaction will also be influenced by the general level of wages in their community. A worker in a job that was once satisfactory may gradually become dissatisfied if incomes in general rise but his income does not. Each generation will tend to base its standards on those of its parents.

Let us now return to the problem of the employer. We cannot describe his situation in terms of a static supply schedule relating numbers of workers to wages. It is not the number of workers which is a function of the wages he pays but the rates of inflow and outflow of workers. The wages he pays will influence the rate at which the workers already employed by him will leave and the rate of flow to him of new workers, workers laid off by other employers, and workers who have voluntarily left other employers.

Most of the workers employed by other employers are satisfied with their jobs and will not be induced to leave because another employer raises wages. His action can tempt away only a few workers who were on the border line between leaving and staying with their employers.[7] In the same way his action may dissuade a small proportion of his own workers from leaving. For the rest his action will tend to bring him a somewhat higher proportion of the floating labor force. The distribution of incomes which are satisfactory among workers will obviously be closely related to the existing distribution of wage incomes. Consequently the higher the wage, relative to the average, offered by an employer, the greater proportion of work seekers who will take his job if they make contact with him. If workers have any information at all about wages, the high-wage employer will also make contact with work seekers sooner, on the average, than others. The rate at which he can hire new workers is therefore increased on both counts.

In markets in which there is usually not very much unemployment an employer who pays relatively high wages will be able, in the long run, to obtain a high-quality work force and have a lower turnover rate than employers who pay lower wages. Moreover, the gain in quality which an employer can achieve by keeping wages above the average by a given amount (percentage) will vary inversely with the rate at which his demand for labor grows relative to growth of labor supply within the community. In those markets in which employers expect a rough balance between supply and demand over fairly long periods, then, there will be a tendency for employers who can get the most out of high-quality labor to pay relatively high wages.

The relative wage required to maintain a labor force of given quality rises as the rate of growth of labor demand grows relative

[7] Lester, *Adjustments to Labor Shortages*, pp. 66–75.

to labor supply, and if the growth of demand exceeds the growth of supply, there will be a minimum relative wage for obtaining any labor at all. In a local labor market, with wages elsewhere given, the resulting competition will drive up the absolute wages until the excess demand is squeezed out. But a general labor shortage over the whole country will not be eliminated so simply unless the money supply is limited (see below, section VII B).

Those considerations apply to the long run. Given an employers' long-run policy, a short-run increase in demand relative to supply offers a very small incentive to increase wages. Over a short period, i.e., a year or two, an employer cannot expect to increase his labor force by more than a couple of per cent by offering a wage increase large enough to be noticed by potential workers. The marginal cost of additional labor during this short period is therefore likely to be several times the wage rate. Moreover, since it is difficult to reduce wages, the employer will be stuck with the wage increase for a long time after the need for it has passed. When the shortage of labor is believed to be temporary, therefore, it is not likely to pay anyone to raise wages in order to attract more labor.

There will, however, tend to be a correlation between changes in employer wage policies and variation in the amount of unemployment. Because of technical change and variation in the rate of growth of different industries and firms, the relative-wage position appropriate to the firms in a market will change from time to time. Since it is difficult and costly to reduce wages, the change in relative wages must be accomplished by firms which wish to improve their positions, raising wages while those who are prepared to lower relative wages take no action. During periods when labor markets are relatively easy, the firms whose rate of growth has increased will have no incentive to change wages, since they have no recruitment difficulties. As soon as labor begins to be scarce, however, the rapidly growing firm will find difficulty in recruiting good-quality labor and will have reason to reexamine its wage policy. If it anticipates recurrence of recruitment difficulties, it will pay the growing firm to raise wages and, since its declining rivals will not reduce wages, average wages will rise.

There will therefore be a certain amount of wage pressure arising from recruiting difficulties when there is excess demand (or relatively low unemployment) even though it arises from long-run rather than from short-run considerations. In practice, of course, the considerations mentioned do not operate by themselves but help to tilt the scale in favor of wage increases in response to other sorts of pressures.

C. Excess Demand and Internal Wage Pressures

Some wage pressure is generated by the recruiting difficulties of employers in periods when demand for labor is high relative to supply. As much or more pressure may be brought to bear on wages by processes within the plant which operate when labor markets are tight.

Tight labor markets from an employer's point of view are easy labor markets from a worker's point of view in the sense that it is easy for a worker to get another job if he leaves his present one. There are always bound to be some workers dissatisfied with their present jobs—because they took unsatisfactory jobs for lack of better ones, because working conditions proved to be less satisfactory than expected, because of changes in job content or personal relations. When jobs are hard to get, workers will tend to stay where they are unless they are very dissatisfied. As soon as it becomes easy to get work, dissatisfied workers will tend to leave, and turnover rates therefore rise. Labor turnover is expensive to employers not only because it creates vacancies but because hiring and breaking in workers itself is costly.

Even workers who are not so dissatisfied as to be prepared to leave always have many sources of dissatisfaction. There are always disputes over job content, promotion and seniority, piece rates, the powers of supervisors, relative rates for different jobs, and so on. Dissatisfaction from these sources is expressed by various forms of non-cooperation which are costly to the employer. Unless workers are very well protected by union rules, the amount of such non-cooperation will be much greater in periods when jobs are easy to get than when they are hard to get. Consequently the pressure on the employer to do something about these difficulties will be greater when he is faced with a tight labor market than when he is not.

The connection of such a situation with wage increases is twofold. (1) Some sources of dissatisfaction are wage-connected, e.g., disputes over piece rates and relative rates on different jobs.[8] (2) Dissatisfaction over specific aspects of jobs interacts with dissatisfaction arising from cost-of-living increases which often accompany tight labor markets or generalized upward pressure on income standards resulting from wage increases elsewhere. The two sorts of dissatisfaction interact, so that wage dissatisfaction is intensified by job dissatisfaction and vice versa. When an employer gives in to a wage demand he therefore not only eliminates the wage dissatisfaction but also reduces the intensity of other types of dissatisfaction. This applies a fortiori to those types of job dissatisfaction specifically associated with wages.

When a trade union is the dominant force in wage negotiations, the situation is obviously somewhat different. The trade union can use the strike as well as more generalized non-cooperation as a weapon for obtaining wage increases. Even without a threat to strike, the union can and does focus the workers' dissatisfaction on its demands and uses these demands as a lever to pry wage increases out of employers. A trade union which never struck or even threatened to strike over wages would still have a powerful influence on wages.

However, wage increases obtained by trade unions should be less closely correlated with excess demand for labor as such (i.e., directly rather than through other variables correlated with demand for labor) than those obtained without trade unions. To the extent that they provide job security and make the expression of job dissatisfaction a group rather than an individual matter, trade unions weaken the connection between morale and ease of reemployment. Moreover, seniority systems and pension plans raise the cost of leaving an employer when jobs are easy to get, so that ease of employment may not have so great an effect on labor turnover as it

[8] Martin Segal, "Factors in Wage Adjustments to Technological Change," *Industrial and Labor Relations Review*, January, 1955, pp. 219–223.

otherwise would. Therefore wage pressure is not so closely related to excess demand for labor as in the non-trade-union case.

However, the ability of the trade union to obtain wage increases and the pressure on it to do so are connected with the employment situation in four other ways.

1. When labor is hard to get, an employer who loses output by strikes or slowdowns cannot expect to make it up by employing more labor even if he has excess capacity. (Of course, it is also quite likely that there will be relatively little excess capacity at times when there is a shortage of labor.)

2. To the extent that trade unions are concerned about the employment of their members, there will be less fear of loss of employment as a result of either technical or product substitution induced by wage increases in periods of high employment than in periods of substantial unemployment.

3. A very large part of the labor force is unorganized, and the variation of demand for labor relative to supply will influence the rate of increase of non-union wages in the manner described above. These changes will influence the bargaining position of trade unions through their effect on the competitive positions of employers and on the relative income position of union workers.

4. As was pointed out above, excess demand for labor gives employers an incentive to raise wages on their own motion. When faced with a trade union, the employer can give in more readily to a wage demand when he has recruiting difficulties than otherwise.

Thus, although trade unions may force wages up even when the labor market is easy, there is every reason to think that, other things equal, they will be more successful in doing so when labor markets are tight.

We can conclude, then, that the rate of increase of money wages will vary with the balance between the supply of and the demand for labor whether trade unions dominate the situation or not. That is not to say that the rate of increase of money wages will be the same in both cases, but we need not enter into that controversial issue.

Moreover, since increases in the cost of living relative to wages are a source of dissatisfaction to workers and the basis of an argument for the trade union which is accepted by employers, the rate of increase of money wages is also influenced by the rate of change of prices.

If there were no trade unions, the influence of the employment situation would make itself felt separately in each local labor market. Rates of unemployment would differ as between local labor markets because of differences in the rates of growth of the industries located in those markets and, to some extent, because of differences in rates of natural increase. As a result, wages would tend to rise faster than the average in those areas blessed with low unemployment and more slowly in those with high unemployment. The resulting tendency toward dispersion would be checked (though not completely) by intra- and interindustry competition and by migration of workers and plants.

Migration of workers is not, of course, primarily determined by wage differentials. It is determined largely by employment opportunity, and therefore workers tend to leave the areas with low employment and slowly increasing wages for those in the opposite situation. As a result the tendency for wages to rise in the high-employment areas is weakened, and conversely. Migration of this type tends to limit inter-area differentials but does not tend to eliminate them.

Since some areas tend to be chronically low-employment low-wage areas, any tendency for demand for labor to grow faster than supply is likely to speed up migration from low-wage areas to high-wage ones and thus to raise the average productivity of the system as well as wages.

It should be noted, of course, that some areas become relatively high-wage areas in one period and remain so for a long time even after they become relatively low-employment areas. Eventually, however, other areas catch up and pass them.

These considerations complicate but do not invalidate the conclusions that the rate of increase of wages varies with the average amount of unemployment. They do suggest, however, that the average wage increase generated by a given average rate of unemployment will vary with the geographical distribution of unemployment and job vacancies.

From the above argument we can conclude that (1) the average rate of increase of wages will tend to vary inversely with the average amount of unemployment or with the average difference between the number of unfilled jobs and the number of unemployed workers; (2) the average rate of wage increase resulting from a given average employment situation will vary with the geographical and industrial distribution of job vacancies and unemployed workers; (3) wages tend to rise faster, *cet. par.*, the higher the rate of increase of cost of living and the higher the rate of profits; (4) when labor markets are tight there is a tendency for workers to drift from low-wage plants and areas to high-wage plants and areas, but that movement takes place at a relatively slow rate.

II. ADJUSTMENT OF LABOR SUPPLY TO DEMAND

We must turn to the question of the relation between the rate of growth of demand for labor and the rate of growth of labor supply.

The effective labor supply at any moment depends on four factors: (1) the size of the population and its composition by age, sex, ethnic origin, and geographical location; (2) labor-force participation rates for each group of the population; (3) the occupational distribution of the labor force (which influences productivity); and (4) working hours.[9]

Each of these factors is constantly changing, and a very large part of the change must be attributed to factors which are largely independent of the balance between supply of and demand for labor.

The growth of population from natural increase depends on social and demographic factors whose relation to economic events is very

[9] Clarence D. Long, "Impact of Effective Demand on the Labor Supply," *American Economic Association, Papers and Proceedings*, 1952, p. 458.

Clarence D. Long, "The Labor Force and Economic Changes," in R. A. Lester and Joseph Shister (ed.), *Insights into Labor Issues*, The Macmillan Company, New York, 1948.

slight. Population growth from immigration, when not controlled by legislation, was much influenced by conditions abroad and by steamship rates, by the level of real-wage rates here, and by the availability of free land. The latter two factors have something to do with the supply and demand for labor, but in rather special ways. The level of real wages is, no doubt, influenced by the balance between supply and demand for labor, but only by a complicated long-run process in which supply and demand for labor are only one element. When immigrants take up free land, they do not contribute either positively or negatively to the balance of supply and demand in the labor market.

Similarly, labor-force participation rates are largely determined by social attitudes, by techniques of production, and by economic factors which are only indirectly related to the balance of supply and demand in the labor market, e.g., the effect of the level of real income on the school-leaving and retirement ages.

The occupational distribution of available employment depends primarily on techniques of production and the structure of demand.

Finally, working hours per year are influenced by social factors and by the level of real income.

Since the underlying factors determining changes in population, labor-force participation, occupational distribution, and working hours vary a great deal over time, it is obvious that a very considerable variance in the supply of labor can be explained by factors which are largely independent of the balance between the supply and demand for labor.

Nevertheless, it can be argued that each of the supply-determining variables is also influenced to an appreciable extent by the balance between supply and demand for labor. These influences work through migration, occupational distribution, and labor-force participation rates, and are more or less long-run in character. In the short-run adjustments through overtime and variation in unemployment are available (without any inconsistency with the observed long-run balance between employment and labor-force growth).

Because of changes in conditions affecting migration and changes in the occupational distribution of the labor force, the effectiveness of the different adjustments is different at different times, so that, when one weakens, a greater pressure is put on the others. In the years before World War I the rate of increase of the labor force of the United States was increased by a very large volume of immigration from abroad. Many of the immigrants would have come regardless of the balance between demand for and supply of labor in the United States. This hard core of immigration included, on the one hand, those who were prepared to take up free land on the frontier immediately, and, on the other, those who were driven from their former homes by intolerably bad political and economic conditions.

From an analytic standpoint, the hard core of immigration had much the same characteristics as the population increase resulting from natural increase, i.e., its contribution to the increase in the labor supply was independent of the movement of demand for labor (except that part of it might have gone elsewhere). But a very large number of potential immigrants were prepared to come to America only if they had a reasonable expectation of finding jobs fairly quickly. When the demand for labor increased more rapidly than the supply from natural increase and from the hard core of immigration, jobs were easy to find and the rate of immigration increased; when unemployment appeared in the United States, immigration fell off. Information on job opportunities reached potential immigrants in Europe slowly and inaccurately, so that the adjustment was far from perfect. Nevertheless, the fact that immigration in the 1890s fell off by a third from the rate reached in the previous decade can hardly have been due to chance, and the same thing is true of the sharp downward adjustment of immigration after the crisis of 1907.[10]

Both during and after the period of heavy immigration from abroad, internal migration had provided a similar flexibility in the effective labor force. In many sections of the country the agricultural population has had a substantial natural increase. The excess population has tended to migrate to urban areas (though, of course,

[10] Thomas Brinley, *Migration and Economic Growth*, Cambridge University Press, Cambridge, England, 1954.

before 1890 some people went to take up farms in newly developed regions) rather than to break up farms or press down wages for agricultural labor. They have not been willing to move, however, unless assured of job opportunities in the towns and cities.[11] As in the case of migration from abroad, farm-to-city migration has taken place chiefly in periods when demand for labor in cities was growing faster than supply. This movement does not, of course, constitute an increase in the total supply of labor in terms of numbers. It does amount to an increase in the effective supply of labor because of the differences between the marginal productivity of additional workers on farms and in industry (when industry is expanding rapidly in scale).

A third group of adjustments involves movement from low-productivity industries to higher-productivity ones, but may not involve any geographical movement.

It was shown in the last section that, as a result of imperfection in the labor market, there are important and persisting differences among wage rates paid to similar labor in different areas and industries. It was also shown that when general excess demand for labor appears, there is a tendency for labor to move from low-wage industries to higher-wage industries. The labor shortage will therefore tend to be more intense for low-wage than for high-wage firms. This may have any of several results: (1) Low-wage firms may raise their rates relative to the others; (2) low-wage firms may reduce hiring standards or improve conditions so as to bring in marginal workers; (3) they may hold down their production; and (4) they may introduce laborsaving devices. The first adjustment may lead to a relative price increase for the firms in question and then lead to consumer substitution away from products of the low-wage firms. Alternatively it may lead to the introduction of laborsaving devices because the price of labor has risen relatively to the price of capital goods.

Which of those adjustments will operate depends on a variety of circumstances, but, given the premise that labor tends to drift

[11] Carter Goodrich, *Migration and Economic Opportunity*, University of Pennsylvania, Philadelphia, 1936, chap. 9.

away from low-wage industries, some of them must operate (unless it just happens that consumer demand increases faster for high-wage than for low-wage industries, in which case the same result is achieved), and the effect will be an increase in the average productivity of the labor force in addition to any other factors tending to make it increase.

A final type of adjustment in the labor force involves long-run changes in the "normal" rates of labor-market participation. At any one moment there are some workers who are definitely in the labor market regardless of conditions in the labor market. There are others who will be in the labor force when wages are falling and job opportunities low who will go out when conditions are more favorable. Still others will tend to act in the opposite way. The proportion of each age, sex, and cultural group in the labor market is, in the short run, a function of job opportunities and wage levels or rates of change in wages. For some groups it will be a rising function of job opportunities, for others a falling one. Over time, cultural changes and changes in the technique of production will shift these functions. A movement along one of these curves is equivalent to increasing or decreasing the number of some type of marginal workers in the labor force. A shift in the curve involves a change in the number of persons in an age-sex group who are marginal workers and a change in the conditions under which they will enter or leave the labor force.

If the shifts in the participation functions were independent of the movements along the curve, there would be nothing more to be said. In fact, however, it seems reasonable to suppose that the shifts are speeded up by movements along the curve. To take the most important concrete case, it is undoubtedly true that the increase in the labor-market participation of women has been due to a series of cultural and technical changes. But those changes did not take place independently of the actual changes in the employment of women.

It was cultural change which first made it socially possible for middle-class young women to try to get jobs, i.e., to come into the class of marginal workers. But if at the time there had been a very easy labor market, they would not have found jobs. Employers would not have made any adaptations in their methods in order to employ women. The example of the pioneers would not have been effective in lowering the social barriers against employment for middle-class women. The changes in social attitudes which we have observed would have taken place to a much smaller extent if labor markets had not been fairly tight during most of the period 1900 to 1930 and again since 1942. Consequently, shifts in labor-market participation functions are a long-run adaptation to excess demand for labor, just as movements along those curves are short-run responses to excess demand for labor. Of course, the argument can be worked the other way, but in the last fifty years or so, at any rate, the adjustments seem to have gone in a direction tending to increase the labor supply when demand is excessive.

To sum up, then, the labor force and its productivity are influenced by excess demand for labor in several ways. The effective labor force changes by (1) overtime, (2) short-run movements along labor-market participation curves, (3) induced long-run shifts in labor-market participation curves, (4) induced immigration, (5) induced farm-city migration. The average productivity of the labor force can be increased by (1) redistribution of workers from low-wage–low-productivity industries to high-wage–high-productivity industries, (2) shift of investment toward laborsaving devices as a result of manpower shortage (including development of new devices), (3) increased use of laborsaving devices in low-wage industries as a result of an increase in wages relative to the prices of capital goods.

At any point in time there is a certain maximum rate of increase in the effective labor force. That maximum will, of course, vary with the length of time over which it is maintained. The effective labor force might be expanded by 4 or 5 per cent per year for one year, but over a ten-year period it might not be possible to obtain an average expansion of more than say, 2 per cent per year. In any case, the maximum rate of expansion of the effective labor force is substantially higher than the rate of expansion with fixed labor-force participation rates, a fixed occupational distribution, and minimum immigration. In general, increase in labor-force participation rates or migration from low-productivity to high-productivity industries will take place only when labor markets are tight and wages are rising.

If the demand for labor grows at a rate less than the rate of growth of supply with fixed participation rates and occupational distribution, participation rates may fall, and agriculture and domestic service may absorb workers rather than release them.

In our discussion of the equilibrium rate of growth we argued that economic systems of the type with which we are familiar are not likely to produce very high equilibrium rates of growth of income. If productivity increases at, say, 2 per cent per year and population increases at 1 per cent per year (without induced immigration), the chance that demand for labor will grow much faster than supply is not very great. In the past, at least, there has been a great deal of play in the rate of growth of the labor force and its productivity, which helped to reduce the frequency of labor shortages.

Labor shortages have occurred, of course, but there seems to be no reason to suppose that actual rate of growth of income has had to fall below the unconstrained equilibrium rate of growth of income very often or for very long on account of labor shortages. It seems likely that really severe labor shortages will occur only when some special transitory stimulus is driving income. Labor shortages which occur without any special temporary stimulus can probably be handled by the mechanisms outlined in Chapter 11. But when demand rises rapidly under the impact of a postwar boom or some other special stimulus, labor shortages will have more serious consequences. The interactions between the movements of demand, wages, and prices under tight labor-market conditions are discussed in the next chapter.

CHAPTER 14

Money Supply, Interest Rates, and Availability of Credit

In the neoclassical literature it was assumed that (1) demand for money to hold is inelastic to the interest rate, (2) the rate of investment is elastic to the interest rate, and (3) wages and prices are flexible. On those assumptions, full employment is guaranteed in the long run, and in the long run the money supply influences the price level but not the real system. The influence of money on the real system was supposed to be confined to transitional periods when the supply of money changes at a higher or lower rate than full-employment real output (or when a crisis causes a shift in the demand for money). In these circumstances the market rate of interest deviates from the real rate, and the distribution of expenditure between consumption and investment is made to deviate from its equilibrium position. At the same time there is either inflation or deflation.

All three of the neoclassical postulates have been challenged in the Keynesian and neo-Keynesian literature. Keynes argued that there is substantial elasticity in the demand for money, and many writers have argued that investment is insensitive to changes in the interest rate.[1] While no one argues that the money supply is completely unimportant, there has been a marked tendency to argue that money supply has relatively little influence on either the real system or the price level. To put it another way, many economists seem to take the view that the price level and volume of output can vary over a wide range with a given stock of money without bringing into play any strong forces tending to drive the volume of money transactions to a fixed level.[2]

[1] J. C. Gilbert, "The Demand for Money: The Development of an Economic Concept," *Journal of Political Economy*, vol. 61, 1953.

[2] W. J. Fellner, "Monetary Policy and the Elasticity of Liquidity Functions,"

It will be argued here, however, that the rate of growth of money income is substantially influenced by the rate of growth of money supply even if it is true that the demand for money is elastic to the interest rate and the demand for investment inelastic. That is so because the interest rate cannot be moved upward very rapidly without causing a temporary increase in the intensity of credit rationing which holds down the rate of investment and reduces the rate of growth of income.

I. INTEREST RATES AND THE DEMAND FOR MONEY

There are three reasons for believing that the demand for money is elastic to the interest rate. Of these, the first is the one employed by Keynes. Keynes maintained that potential security buyers always have in view a normal level of rates, so that a rise in the actual rate above the normal rate leads to an expectation of a future fall and conversely. Those who buy or hold now will obtain a capital gain in future if the rate falls and vice versa. In this view, the attractiveness of holding securities becomes greater as the yield rises.

On a priori grounds there is no reason why the argument should not be turned just the other way. In many fields, trend projection seems to be the dominant influence on expectation. It would not, therefore, be surprising if it turned out that a rise in rates led to an expectation of a further rise and vice versa. In that case the speculative motive would work in the opposite direction to the one postulated by Keynes. It is almost certainly true that most persons who take an interest in security prices will be influenced by both types of consideration, and that the balance between the two considerations will depend on the circumstances surrounding the rate movement in question. There is therefore no strong a priori presumption in favor

---

Review of Economics and Statistics, vol. 30, 1948; and *Monetary Policy and Full Employment*, University of California Press, Berkeley, 1946, pp. 137–235.

James Tobin, "Liquidity Preference and Monetary Policy," *Review of Economics and Statistics*, vol. 29, 1947; and "Monetary Policy and the Management of the Public Debt: The Patman Inquiry," *Review of Economics and Statistics*, vol. 35, 1953.

of Keynes's "normal" hypothesis, but the same conclusion may be reached by a different route.

A. Risk Aversion and Uncertainty

Any holder of long-term debt has subjected himself to a risk of capital loss resulting from changes in the rate of interest. Even if everyone considers it just as likely that yields will rise as that they will fall, many persons will require a return for undergoing this risk. Some persons are more willing to take a given risk than others, and there will be differences in the variance of subjective probability distributions of future yields. Therefore the amount of money which the community as a whole will wish to hold will vary inversely with the level of yields (and with the structure of yields of different maturities). Thus, even if no one has any stronger belief in a rise than in a fall in rates, there should be some elasticity in the supply of loanable funds with respect to the interest rate. However, that consideration operates only on those asset holders who require the use of their funds at a date earlier than the date of the maturity of the securities they buy. An asset holder who considers it very improbable that he will wish to sell out his portfolio in less than twenty years (a university or an insurance company with growing assets) can always buy securities to hold to maturity. He takes no risk of capital loss, though he does take the risk that in order to avoid a capital loss he may have to give up more favorable opportunities in the future. That is not the sort of risk to which risk aversion should apply. It is presumably that consideration which keeps institutional cash holdings down to minimal levels.[3]

Financial intermediaries who pool liquidation requirements and are therefore able to hold a large proportion of their assets without fear of liquidation at an unfavorable time help to remove the effect of risk aversion from individuals and businesses. The existence of bonds of the United States Savings Bonds Series E type has the same effect.

[3] O'Leary, "The Institutional Savings Investment Process and Current Economic Theory," *Proceedings of the American Economic Association, American Economic Review,* March, 1954, pp. 455–470.

It is doubtful, therefore, whether the cash holdings of households or institutions are very sensitive (on speculative grounds) to differences in the level of interest rates which persist for very long. However, if rates change very quickly, uncertainty as to future levels of rates may increase, and potential security buyers may refrain from making new commitments for a time. The significance of that consideration is discussed below. Aside from such dynamic considerations, there is some doubt whether institutions and households really do contribute very much to the static interest elasticity of demand for money.

Business firms, however, do hold large amounts of liquid assets, and they cannot hold long-term securities to maturity without risk. They hold liquid assets either because of temporary variations in cash inflows and outflows or to provide for contingencies which might increase their cash requirements. For that reason, business firms generally tend to hold their liquid assets very largely in the form of short-term debt or cash. However, when the spread between long and short rates is wide, business firms may be induced to place a certain amount of their funds in long-terms.

If firms confine themselves to holding short-term debt of maturities corresponding to known cash requirements, they need not assume any risk of changes in the interest rate. The principal barrier to the full exploitation of the interest is the cost and trouble of managing a rapidly fluctuating portfolio of short-term obligations. Since large sums can be handled almost as easily as small ones, the reward for taking trouble varies with the size of the firm's stock of idle funds. At low interest rates, only very large firms can profit by placing funds on the short-term securities market, while at higher ones, an adequate return can be obtained on a much smaller balance. The cost of managing a short-term portfolio also varies with the number of transactions to be made. The number of transactions required varies with the proportion of the firm's peak liquid assets invested. If the firm holds no cash, it will have to make thousands of transactions per year. On the other hand, if it invests only funds which are held against remote contingencies, the number of transactions will be virtually determined by the frequency of maturities.

At high rates of interest, therefore, it will pay a firm to hold its cash to a minimum and engage in a large number of security transactions, and conversely at low rates.[4]

The tendency for transactions balances to vary with the interest rate may be at least as important as the speculative considerations emphasized by Keynes. Essentially the same argument applies to the transactions balances of households, with savings deposits playing the role played by short-term debt for business.

B. Empirical Evidence on Interest Elasticity of Demand for Money

There are a number of pieces of empirical evidence which can be used to test various hypotheses about the relation between the demand for money and the level of interest rates. Most of this evidence has been generated by the movements of interest rates and money supplies in the years since the late 1920s. From 1900 to 1929 interest rates moved within a fairly narrow range, and the circuit velocity of demand deposits also moved within a narrow range. The experience of that period is not very useful from a hypothesis-testing point of view.

The experience of the 1930s appears to confirm the Keynesian hypothesis. The fall in transactions and the inflow of gold generated an enormous volume of excess reserves. Short-term interest rates on government securities fell to very low levels, but long-term rates fell slowly and never dropped below about 2 per cent. Any individual commercial bank could have increased its earnings by buying additional governments. If all banks had tried to do that, they would have driven the rates on long-term governments down almost to zero. There seems to be no reason why banks should not have attempted to increase their bond portfolios to a greater extent than they did except their fear that rates might rise at a time when

[4] James Tobin, "The Interest Elasticity of Transactions Demand for Cash," *Review of Economics and Statistics*, August, 1956, pp. 241–247.

N. J. Baumol, "The Transactions Demand for Cash: An Inventory Theoretic Approach," *Quarterly Journal of Economics*, vol. 66, 1952.

liquidation would be necessary. That argument is consistent with both the Keynesian return-to-normal hypothesis and the unbiased uncertainty hypothesis. However, the circumstances were abnormal. After the wave of bank failures in the early 1930s, banks undoubtedly had a greater fear of a liquidity crisis than ever before. Each bank therefore wished to maintain a high degree of liquidity, and at the same time feared that a general struggle for liquidity would depress bond prices at the very moment at which sales would be necessary. In these circumstances the market risk on bonds kept their yields from falling below 2 per cent in spite of the great volume of excess reserves. But one cannot judge the importance of market risk as a determinant of bond yields from the experience of the 1930s.

Some studies of business cash holdings show a marked relation between the movement of interest rates and business demand for cash (relative to transactions).[5] This amounts to saying that business cash holdings rose relative to transactions during the period when interest rates were falling and remained relatively high thereafter. That evidence is consistent with the transactions-demand argument given above. It can be partially explained away. The rise in cash holdings during the 1930s can be explained in terms of fears of a liquidity crisis. That argument can hardly be used to explain the postwar situation.

In the period since 1950, interest rates have risen, and the velocity of circulation has risen substantially. Between 1950 and 1955 the velocity of circulation rose by about 25 per cent. A large part of the increase was due to a decline in corporate cash holdings relative to GNP. Between 1950 and 1955 corporate cash holdings rose by 12 per cent, while GNP rose by 50 per cent. Moreover, business hold-

[5] A. Kisselgoff, "Liquidity Preference of Large Manufacturing Corporations (1921–39)," *Econometrica*, vol. 13, 1945, pp. 334–344.

J. N. Behrmann, "The Short Term Interest Rate and the Velocity of Circulation," *Econometrica*, vol. 61, 1948, pp. 185–190.

Henry Allen Latane, "Cash Balances and the Interest Rate—A Pragmatic Approach," *Review of Economics and Statistics*, vol. 36, 1954, pp. 456–460.

ings of cash and government securities combined rose by only 14 per cent in the same period.

However, the upward movement of velocity and the downward trend in corporate liquidity were not very closely correlated with the short-term movements of interest rates. It may be reasonable to suppose that after a long period of low interest rates it takes business firms some time to adjust their cash-management practices to the new situation. The full effect of an increase in interest rates is felt only after a fairly long-distributed lag. Then if the trend of rates is upward, the trend of velocity is upward, and velocity need not show much response to small short-term movements in rates. In addition, there must be some lag in the response to downward movements of rates. Once a corporation has made the organizational changes necessary to bring about economy in the use of cash, it cannot quickly disband its organization in response to small downward movements in interest rates.

We might, therefore, regard the period 1950 to 1956 as a period of adjustment to higher interest rates, in which excess liquidity has been gradually squeezed out of the economy. It remains to be seen how much further velocity can be raised by further increases in rates.

It must be noted, however, that other forces were at work in raising velocity during the postwar period. Business and personal liquidity were excessive by any standard at the end of World War II. Moreover, increasing confidence makes people willing to operate with less liquidity. It is possible to explain away the whole movement of velocity since 1930 by reference to changes in confidence.

The Keynesian liquidity-preference argument, therefore, achieves only dubious support from the empirical evidence, but the transactions-cost argument receives fairly strong support.

Since there are both theoretical and empirical grounds for expecting that the ratio of cash holdings to transactions will vary with the interest rate, we can tentatively suppose that it is possible for the volume of transactions associated with a given money stock to vary over a fairly wide range.

C. Rationing

It does not follow that the rate of growth of money supply has no influence on the rate of growth of aggregate income. Suppose we start with some given set of credit conditions, that is, terms and conditions on which loans can be obtained, and quality standards which are applied by lenders. Suppose that under those credit conditions the rate of growth of money income exceeds the rate of growth of money supply. Interest rates will tend to rise. According to the usual theory of the demand for money, the increase in interest rates tends to reduce the demand for money (relative to transactions), thus permitting expenditure to grow faster than supply.

In fact, however, mortgage rates and commercial-bank customer's loan rates will not rise immediately in response to excess demand for funds. Like any other administered price, mortgage rates and customer's rates are changed infrequently.

A lender faced with a demand for funds in excess of his lending capacity who does not raise his rates at once has two alternatives. He can sell open-market securities, or he can tighten credit terms, i.e., increase required balances, refuse borrowers of marginal credit standing, require higher down payments on mortgages, and so on. He is likely in fact to do both. A tightening of credit standards, even if it is only temporary, will slow down the rate of growth of income by reducing investment.

Any particular lender can sell open-market securities to other financial institutions. But if lenders generally are faced with an excess demand for loans, they can obtain the power to make additional loans only by selling securities to holders of cash, i.e., to businesses or individuals. If businesses and individuals are satisfied with their portfolios at the existing rates, lending institutions can obtain lendable funds only if interest rates rise. The resulting fall in the price of long-term securities may induce an expectation of further decline and will almost certainly cause an increase in uncertainty about future interest rates. That will induce a greater fall in prices than would have taken place with fixed expectations. In an uncertain market with falling bond prices, institutional investors may tend to postpone commitments on private placements, and investment banking houses will hesitate to underwrite new issues because of the difficulty of making a price which will ensure quick sale.[6]

The weakness in the new-issue market can be removed only by a cessation of the rise in rates or by a fall. That will be brought about by the reduction in the rate of investment caused by the weakness of the new-issue market. It may, however, take some time for the effect of a reduction in new issues to be felt because of the lagged effect of previous issues. When the effect is felt, the whole process will reverse itself. A rise in bond prices will take place which will be reinforced by speculative effects. How far down rates will go depends on the speed of response of investment to the improvement in money-market conditions and on the length of time between the reduction in new issues and the recovery of bond prices. The latter consideration is important because capital will have grown faster than income during the downswing, so that when easy money begins to rule, the system will still tend to grow at a slower rate than the equilibrium. If the adjustment process takes a long time and the response lag of investors is greater on the upswing than on the downswing, interest rates may fall to the initial level before they begin to rise again, and there may therefore be no net upward movement in rates. In any case, the net upward movement across the cycle will be much smaller than the upswing taken by itself.

It is therefore possible to hold the rate of growth of aggregate demand down to the rate of growth of money supply without any trend in the interest rate or at any rate with a very slow upward-trend movement. The control is provided by an appropriate frequency of reduction of growth of income below its "easy-money" equilibrium level.

It seems reasonable to suppose that a sufficiently rapid increase in interest rates would bring about the types of rationing described above. It does not follow that small changes in rates will have much effect. If the level of rates is raised by a quarter of a point every three

[6] Robert Roosa, "Interest Rates and the Central Bank," in L. H. Dupriez et al., *Money, Trade, and Economic Growth Essays in Honor of John H. Williams*, The Macmillan Company, 1951.

months, no one may be very much frightened, and the lag in the adjustment of "administered" rates may have little effect.

The history of the National Banking system shows that one can check the growth of income by making money sufficiently tight (intentionally or unintentionally), but whether less drastic measures will be sufficient is still in doubt.

The tight-money policy of 1955–1956 does seem to have had some effect on residential building (though how much of the decline in building was due to the monetary situation is still in doubt). Much of the effect was due to the rigidity of the rates on guaranteed mortgages. It is alleged in the press that small business has had difficulty in obtaining bank loans. Elsewhere the effects of the tight-money policy are less apparent. High rates are said to have discouraged some local governments from borrowing. There is no indication that institutional lenders are holding extra cash or refusing to switch out of governments into private obligations. The new-issue market has shown only occasional weakness.

The Federal Reserve Board could have followed a more drastic policy, and at some point its policy would undoubtedly become effective. It does seem likely that a sufficiently sharp upward movement of rates might cause a sharp reduction in investment without causing a major depression. However, that is a judgment which cannot be supported by the facts.

II. MONEY SUPPLY, WAGES, PRICES, AND GROWTH

We have now sketched out very roughly the structure of labor markets and money markets. The problem now is to see how these two sets of markets interact with the growth of real income and with one another. We can proceed by sketching out a number of alternative situations with respect to the relative rates of growth of aggregate demand, labor supply and money supply.

In our discussion of the movements of real income in Chapters 10 to 12 we assumed that the parameters of the income-generation system were given without regard to labor-market conditions and money-market conditions. We now have to take greater care about the relation between labor-market conditions, money-market conditions, and the parameters of the investment equations. The simplest procedure seems to be to define standard "easy-labor-market," "easy-money-market" conditions. If the amount of labor available exceeds the amount demanded by more than a certain amount, labor markets will be as "easy" as they can get. No firm's output will be held down by scarcity of labor, and no firm will fail to invest because of lack of man power. Further increases in labor supply relative to demand will not increase output or investment. A higher level of demand for labor relative to supply will reduce the amount of unemployment, but for the unemployment to be reduced a certain number of unfilled vacancies has to exist. However, up to a point, that makes very little difference, because most firms will be able to get along with a slightly less than optimum labor force without any significant reductions in output (though at higher money cost if they resort to overtime operations or use obsolete plant). But when average unemployment falls to less than 3 per cent, more serious labor shortages will develop. Investment may be discouraged, and the growth of output may be determined by the growth of the labor force rather than by the expenditure equations. Let us call the rate of growth of real income which would rule if there were more than 3 per cent unemployment the "unconstrained" rate of growth of demand.

The rate of growth of effective labor supply is also a function of the difference between the demand for and supply of labor. When there is a substantial excess supply of labor, the rate of growth of the labor supply will be determined by natural increase of population, the hard core of immigration, the distribution of demand, and by techniques of production. Let us call the rate of growth of labor supply which would occur under those conditions the "no-pressure" rate of growth of labor supply. At some level of excess demand for labor, jobs become fairly easy to get, and the rate of growth of labor supply will rise above the no-pressure rate.

Now suppose that the unconstrained rate of growth of demand for labor exceeds the no-pressure rate of growth of labor supply. Suppose that at the outset there is a certain amount of unemployment.

After a time the unemployment will be reduced to less than 3 per cent. As a result, there will be unfilled jobs, and the rate of growth of real income will fall below the unconstrained rate. This will induce an inventory cycle of the type described in Chapter 11. Meanwhile, however, tight labor markets will have caused labor supply to increase at a rate higher than the no-pressure one. This process will repeat itself as long as the same unconstrained rate of growth of demand and no-pressure rate of growth of supply continues to rule. As a result the average rate of growth of real income will be below the unconstrained rate and above the no-pressure rate. As a by-product, money wages will rise during the booms and will not fall by an equal amount during the slumps. But although a wage-price spiral will result, it will proceed at a relatively slow rate and may be offset by productivity increases.

As a second case, suppose that as a result of some temporary situation, such as a postwar boom, the unconstrained rate of growth of demand for labor is well above the no-pressure rate of growth of supply. Then the rate of growth of demand will not fall below the rate of growth of labor supply even when the rate of growth of real demand is held down by labor shortages. Wages will tend to rise rapidly, and there will be a rapid wage-price spiral. Inventory speculation may result, or there may be a general willingness to invest more because of rising prices. Such a boom is bound to be checked for one reason or another. Investment may be checked by limitations on money supply or shortages of capital goods. Alternatively, manpower shortages will eventually depress the returns on investment and offset the effects of rising prices. Without any of those factors coming into play, the boom may be ended by a chance event or by a general feeling that the boom is based entirely on expectations. We have already discussed the depression which results. On the whole, cycles of this sort are likely to occur only in special circumstances, e.g., the postwar boom of 1919–1920.

Monetary constraints operate through different channels but have effects very similar to the effects of labor-supply constraints. Any ratio between the stock of money and the level of money income will produce a certain structure of interest rates and a certain set of standards for the granting of credit. There is some ratio of money supply to money income at which minimum credit standards will be reached. Further increases in the supply of money (relative to money income) will have no effect on investment. Call the rate of growth of money income which results from a monetary situation in which minimum credit standards rule the "unconstrained rate of growth of money income." That rate will reflect the rate of growth of real income and the rate of change of prices produced by the labor-market situation.

If the unconstrained rate of growth of money income is below the rate of growth of money supply, the actual rate of growth of money income will equal the unconstrained rate (or move toward it after any initial tightness of money is removed).

If the unconstrained rate of growth of money income is above the rate of growth of money supply, interest rates will be forced up, and when that happens, credit rationing will become temporarily tighter. Investment will be reduced (by comparison with the no-constraint situation), and the rate of growth of both real and money income will be reduced. An inventory cycle of the type described in Chapter 11 will follow. During the downswing, money will become easier and interest rates will fall (though not necessarily to their starting point). At the same time, because the real rate of growth has been reduced by the monetary constraint, the labor-supply situation will be easier than without the monetary constraint. As a result, wages will not rise so quickly as they would have done without the constraint.

The rate of growth of money income is adjusted to the rate of growth of money supply by (1) reducing the rate of growth of real income, (2) holding down the rate of increase of wages, (3) raising interest rates and thus increasing the volume of transactions which can be handled by a given money supply. Of course, the rate of growth of money supply need not be a given datum. If the monetary authorities wish to avoid causing unemployment, they will increase the money supply as fast as need be to prevent a shortage of credit.

If a speculative boom in stocks is under way when money is relatively easy, the tightening of credit may check the rise in stock prices, and that may cause speculators to sell out and thus lead to a rapid decline in stock prices.

Both monetary and labor shortages may interact with other factors tending to depress income. In other words, a decline in investment is much more likely to occur when labor and money markets are tight than when they are easy. But tight labor and money markets in themselves are not likely to cause major depressions.

CHAPTER 15

Summary and Conclusions

Although the prediction and control of the future course of events is one of the principal ultimate objectives of any scientific endeavor, the first task of any theory is to explain what has already happened. It is to be hoped that the set of hypotheses developed here will prove to be of some use in prediction and policy making in the future, but at the moment they must be regarded as hypotheses which provide an explanation of the movements of real and money income in the United States in the period since the Civil War.

The explanation in question can be summarized in the following way. We can suppose that the movements of aggregate income are determined by a capital-adjustment process in which the income and capital stocks existing at one moment determine the rate of change of income and capital stock, and therefore later values of income and capital stock, with the process being repeated indefinitely. That simple basic notion is complicated in three ways: (1) A whole set of capital stocks is involved, so that the process has to be regarded as a generalization of a dynamic input-output scheme of the type developed by Leontief rather than as a generalization of an aggregative multiplier-accelerator model. (2) There are a number of complicated lags in the process. Some of the lags are associated with the connection between expectations and past movements of income. Others are associated with the structure of markets for residential and commercial buildings and with their construction. (3) The parameters of the capital-adjustment process are subject to change with the passage of time. These parameter changes are due to (*a*) changes in expectations (other than those due to the past history of income movements), (*b*) changes in competitive structure, (*c*) changes in financial institutions, (*d*) changes in relative prices of capital goods and wages, which are in turn due either to differences in rates of technical progress between industries or to differential wage movements, (*e*) changes in the rate of improvement of technique, (*f*) changes in average capital intensity due to changes in technique, (*g*) changes in the distribution of demand among industries, (*h*) speculative activity based on rapid growth of particular industries, (*i*) speculative activity generated by wage-price spirals, (*j*) changes in credit conditions due to speculation in securities, a difference between the rate of growth of demand for and supply of money, or extraneous, e.g., political causes. It will be noted that some of those parameter changes could be dealt with systematically by enlarging the scope of the capital-adjustment process. The choice is essentially one of convenience.

Even when its parameters are fixed, a multisector dynamic process involving both lags and cumulants is almost certain to produce cyclical movements in the variables if it is once disturbed from its equilibrium path (assuming that it has one).

It is argued here that, when the system has an equilibrium position, an important movement in the parameters of the system can cause one major swing in aggregate income, and that the secondary, or induced, cycles following the initial disturbance are likely to damp out rapidly. Minor cycles may therefore be explained without introducing a change in parameters, but not so major cycles. It is possible, however, that under appropriate conditions a major induced cycle could develop as a result of the structure of the housing industry.

Major depressions can be generated in several different ways.

1. The "normal" parameters of the income-generation system, i.e., those which would rule in the absence of speculative or other temporary factors, may be such that the system does not have a real positive equilibrium rate of growth. In that case, income may rise for a time because the initial level of income is high relative to initial capital stocks, but the system tends to run down; i.e., capital tends to accumulate more rapidly than income until income stops rising. There appear to be no examples of that type of depression, though it is possible that an example would have been provided by

the 1920s had there not been so many other good reasons for a depression.

2. The collapse of a speculative boom in a particular industry can bring about a general depression; the depressions of 1873 and 1883 seem to provide good examples. The speculative factor probably also played a role in 1893.

3. The collapse of a speculative boom based on rising prices in a wage-price spiral can also produce a depression. The wage-price spiral itself can result from a tendency for demand for labor to outrun supply. The depression of 1920–1921 is a case in point. It should be noted that, although 1920–1921 seems to be the only case in which a speculative movement based on a wage-price spiral was the principal cause of the depression, such speculative movements may have played a role in other depressions as well.

4. When money income grows faster than the money supply, the growth of income will be at least temporarily checked by credit rationing. That will not ordinarily cause a depression. But if a speculative boom in securities develops before money becomes tight, a stock-market crash will result. Such a crash can cause at least a temporary reduction in the rate of investment. That process is at least a possible explanation for the depression of 1907.

5. A financial panic can develop without tight money either because a stock-market boom collapses "from its own weight," as in 1929, or because of extraneous factors, such as the gold crisis of 1893. A financial panic, whatever its cause, is bound to have some adverse effect on the rate of investment and can bring on a major depression with the aid of other factors.

6. As already noted, a housing boom generated by earlier variations in the growth of aggregate income will be followed by a housing slump, which could bring about a general depression. That possibility becomes more important if the housing boom involves speculative activity in apartment and commercial construction. It was argued in Chapter 7 that instability of construction during the 1920s played a major role in bringing on the depression of 1929, though it had plenty of assistance from other sources.

Given a substantial decline in the rate of investment, the amplitude and duration of the ensuing depression depends on a number of factors governing the speed with which income moves downward under the impact of the initial disturbance. Among the factors, the most important are:

1. The extent to which the secondary decline in expenditure is limited by lags in the responses of (*a*) consumption to income, (*b*) dividends and entrepreneurial withdrawals to profits, (*c*) housing construction to income change.

2. The support given to consumption generally and housing demand in particular by rising population.

3. The extent to which the rapid growth of particular sectors or areas assists in maintaining investment.

4. The sensitivity of the marginal efficiency of investment to changes in demand and the extent to which "capital rationing" has built up a backlog of investment demand during the boom.

5. The response of prices to the change in income and the speculative and asset effects of the decline in prices.

6. The stability of financial institutions.

The amplitude and duration of the depression also depends on the extent to which the initial cause of the depression can be corrected in the face of the general decline in incomes and prices.

The argument given above indicates that the observed "cycles" cannot be accounted for by a mechanism which tends to produce the same sequence of events over and over again. The observed cycles differ not only in detail but also in the basic causal mechanisms behind them, unless we describe that mechanism in the broadest and most general terms. The appearance of regularity in the timing of major depressions arises from two causes:

1. It takes a considerable length of time for the economy to recover from one depression and reach a situation in which any of the processes which bring about depressions can get under way (the short interval between recovery from the depression of the 1870s and the start of the one in the 1880s must be taken as a tribute to the energy and "never-say-die" spirit of American railway promoters).

2. There are enough ways for a depression to start, once a period

of prosperity has continued for a few years, that if one does not work, another will. Even so, the period from 1897 to 1929 was a period in which the major depressions lasted only a little over a year each. The theory developed here does not produce a regular cycle, but there does not seem to be a compelling theoretical reason why it should do so. The pattern of movements of income suggested by the theory developed here, in interaction with the historical changes in the structure of the economy which have occurred, does produce income movements of the types that we have observed. At the same time the hypothesis which underlies the theory of income movements seems to be consistent with the observed facts. It must be admitted that this theoretical structure has not been submitted to very sharp empirical tests at either the micro or the macro level. There is little doubt that the hypotheses developed here will have to be substantially revised when submitted to adequate empirical tests. One can only hope that enough will remain to justify the general approach to the cycle problem used here.

Index