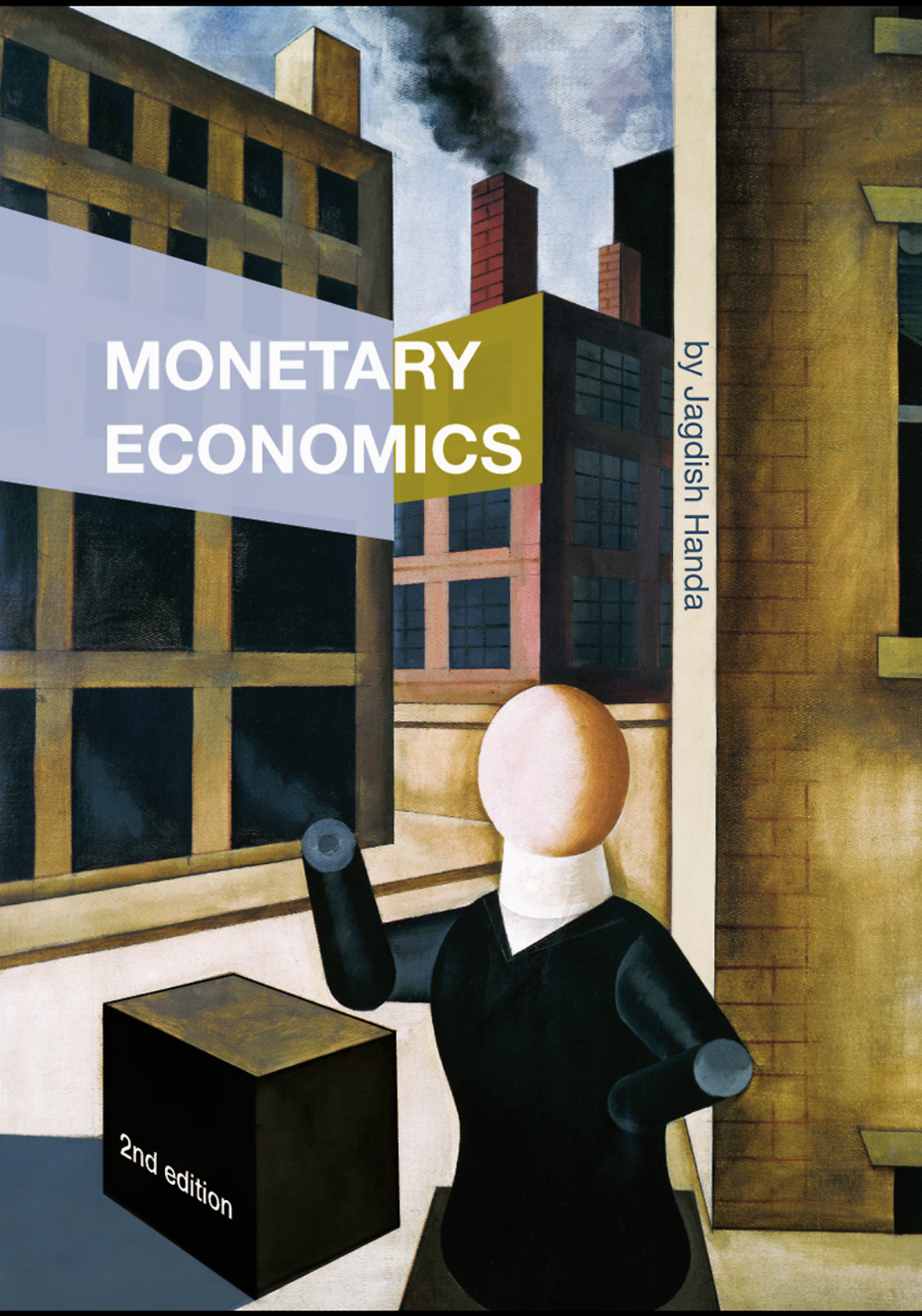
MONETARY ECONOMICS

by Jagdish Handa

2nd edition



Monetary Economics, 2nd Edition

This successful text, now in its second edition, offers the most comprehensive overview of monetary economics and monetary policy currently available. It covers the microeconomic, macroeconomic and monetary policy components of the field. The author also integrates the presentation of monetary theory with its heritage, stylized facts, empirical formulations and econometric tests.

Major features of the new edition include:

- Stylized facts on money demand and supply, and the relationships between monetary policy, inflation, output and unemployment in the economy.
- Theories on money demand and supply, including precautionary and buffer stock models, and monetary aggregation.
- Cross-country comparison of central banking and monetary policy in the US, UK and Canada, as well as consideration of the special features of developing countries.
- Competing macroeconomic models of the Classical and Keynesian paradigms, along with a discussion of their validity and consistency with the stylized facts.
- Monetary growth theory and the distinct roles of money and financial institutions in economic growth in promoting endogenous growth.
- Excellent pedagogical features such as introductions, key concepts, end-of-chapter summaries, and review and discussion questions.

This book will be of interest to teachers and students of monetary economics, money and banking, macroeconomics and monetary policy. Instructors and students will welcome the close integration between current theories, their heritage and their empirical validity.

Jagdish Handa is Professor of Economics at McGill University in Canada and has taught monetary economics and macroeconomics for over forty years.

Monetary Economics, 2nd Edition

Jagdish Handa

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

First published 2000
Second edition published 2009
by Routledge
2 Park Square, Milton Park, Abingdon, Oxon OX14 4RN

Simultaneously published in the USA and Canada
by Routledge
270 Madison Ave, New York, NY 10016

Routledge is an imprint of the Taylor & Francis Group,
an informa business

This edition published in the Taylor & Francis e-Library, 2008.

"To purchase your own copy of this or any of Taylor & Francis or Routledge's collection of thousands of eBooks please go to www.eBookstore.tandf.co.uk."

British Library Cataloguing in Publication Data
A catalogue record for this book is available
from the British Library

Library of Congress Cataloging in Publication Data
A catalog record for this book has been requested

ISBN 0-203-89240-2 Master e-book ISBN

ISBN 10: 0-415-77209-5 (hbk)
ISBN 10: 0-415-77210-9 (pbk)
ISBN 10: 0-203-89240-2 (ebk)

ISBN 13: 978-0-415-77209-9 (hbk)
ISBN 13: 978-0-415-77210-5 (pbk)
ISBN 13: 978-0-203-89240-4 (ebk)

To Sushma, Sunny and Rish

Contents

Preface

This book represents a comprehensive presentation of monetary economics. It integrates the presentation of monetary theory with its heritage, its empirical formulations and their econometric tests. While its main focus is on monetary theory and its empirical tests rather than on the institutional monetary and financial structure of the economy, the latter is brought in wherever needed for elucidating a theory or showing the limitations to its applicability. The illustrations for this purpose, as well as the empirical studies cited, are taken from the United States, Canada and the United Kingdom. The book also elucidates the significant differences between the financially developed economies and the less developed and developing ones.

In addition, the presentation also provides an introduction to the main historical patterns of monetary thought and the diversity of ideas in monetary economics, especially on the effectiveness of monetary policy and the contending schools in monetary theory and policy.

Our presentation of the theoretical aspects of monetary economics is tempered by the goals of empirical relevance and validity, and intuitive understanding. The derivation of the theoretical implications is followed by a discussion of their simplifications and modifications made in the process of econometric testing, as well as a presentation of the empirical findings.

Part I of the book consists of the introduction to monetary economics and its heritage. The latter is not meant to be exhaustive but is intended to illustrate the evolution of monetary thought and to provide the reader with a flavor of the earlier literature on this subject.

Part II places monetary microeconomics in the context of the Walrasian general equilibrium model. To derive the demand for money, it uses the approaches of money in the utility function and in the production function. It then derives the Walrasian results on the neutrality of money and the dichotomy between the monetary and real sectors of the economy.

Part III focuses on the demand for money. Besides the usual treatment of transactions and speculative demands, this part also presents models of the precautionary and buffer stock demand for money. The theoretical chapters on the components of money demand are followed by three chapters on its empirical aspects, including a separate chapter on the criteria and tests underlying monetary aggregation.

Part IV deals with the supply of money and the role of the central bank in determining the money supply and interest rates. It compares the desirability of monetary versus interest rate as operating targets. This part also examines the important policy issues of the potential conflicts among policy makers, central bank independence, time-consistent versus discretionary monetary policies, and the credibility of monetary policy.

No presentation of monetary economics can be complete without adequate coverage of monetary policy and its impact on the macroeconomy. Proper treatment of this topic requires knowledge of the underlying macroeconomic models and their implications for

monetary policy. Part V focuses on money and monetary policy in the macroeconomy. It covers the main macroeconomic models of both the classical and Keynesian paradigms and their monetary implications. This coverage includes extensive analysis of the Taylor rule for targeting inflation and the output gap, and new Keynesian economics.

The remaining parts of the book deal with special topics. Part VI deals with the theories of the rate of interest and of the term structure of interest rates. Part VII presents the overlapping generations models of fiat money and compares their implications and empirical validity with those of the theories based on money in the utility function and money in the production function. Part VIII addresses monetary growth theory, and assesses the contributions of both the quantity of money and those of financial institutions to output growth. To do so, it covers the neoclassical growth theory with money as well as endogenous growth theories with money.

Comparison with the first (2000) edition

This edition has extensive revisions and new material in all its chapters. However, since the major ferment in monetary economics in the past decade has been in monetary policy and monetary macroeconomics, most of the additional material is to be found in the chapters on these issues. Chapter 12 has more extensive discussion of central bank independence, time consistency versus intertemporal re-optimization, and credibility. Chapter 13 is a new chapter on the determination of aggregate demand under the alternative operating targets of money supply and interest rates. Chapter 14, on the classical paradigm, now starts with a presentation of the stylized facts on the relationship between money, inflation and output, and includes more detailed evaluation of the validity of the latest model, the modern classical one, in the classical paradigm. Chapter 15, on the Keynesian paradigm, has considerably more material on the Taylor rule, and on the new Keynesian model, as well as discussion of its validity. Chapter 16, on the role of credit markets in the macroeconomy, is entirely new. Chapter 17 has been expanded to include compact models of the new Keynesian type, in addition to the Lucas–Sargent–Wallace ones of the modern classical variety, as well as including greater discussion of the validity of their implications. Chapter 21, on the overlapping generations models, now starts with a presentation of the stylized facts on money, especially on its demand function, so as to more clearly assess the validity of the implications of such models.

Level and patterns of use of this book

This book is at the level of the advanced undergraduate and graduate courses in monetary economics. It requires that the students have had at least one prior course in macroeconomics and/or money and banking. It also assumes some knowledge of differential calculus and statistics.

Given the large number of topics covered and the number of chapters, this book can be used over one semester on a quite selective basis or over two or three semesters on a fairly complete basis. It also offers considerable scope for the instructors to adapt the material to their specific interests and to the levels of their courses by exercising selectivity in the chapters covered and the sequence of topics.

Some suggested patterns for one-term courses are:

1. Courses on monetary microeconomics (demand and supply of money) and policy: Chapters 1, 2, 3 (optional), 4, 5, 7–12.

2. Courses on monetary macroeconomics: Chapters 1, 2, 13–17 (possibly including Chapters 18–20).
3. Courses on monetary macroeconomics and central bank policies: Chapters 1, 2, 10–19.
4. Courses on advanced topics in monetary economics: Chapters 3, 6, 16–24.

A first course along the lines of 1, 2 or 3 can be followed by a second course based on 4.

McGill offers a tandem set of two one-term graduate courses covering money and banking and monetary economics. The first term of these is also open to senior honours students. This book came out of my lectures in these courses.

My students in the first one of the two courses almost invariably have shown a strong interest in monetary policy and macroeconomics, and want their analyses to be covered at an early stage, while I want also to cover the main material on the demand and supply of money. With two one-semester courses, I am able to allow the students a wide degree of latitude in selecting the pattern in which these topics are covered. The mutually satisfactory combination in many years has often been to do in the first semester the introductory Chapters 1 and 2, monetary macroeconomics (Chapters 13 to 17), determination of interest rates (Chapters 19 and 20) and possibly monetary growth theory (Chapter 24). The second term then covered money demand and supply (Chapters 4 to 10) (excluding Chapter 6 on the precautionary and buffer stock demands for money) and central banking (Chapters 10 to 12). However, we have in some years chosen to study the money demand and supply chapters before the monetary macroeconomics chapters. This arrangement left the more theoretical, advanced or special topics to be slotted along with the other material in one of the terms, or left to another course. The special topics chapters are: 3 (general equilibrium with money), 6 (precautionary and buffer stock models of money demand), 16 (credit markets), 17 (compact macroeconomic models with money), 18 (Walras's law and the interaction among markets), 21 to 23 (overlapping generations models with money) and Chapter 24 (growth theory with money).

Acknowledgments

I am indebted to my students in monetary economics who suffered – and hopefully benefited – from several drafts of this manuscript. Many helped to improve it.

It is, as always, a pleasure to acknowledge the love and support of my wife, Sushma, and sons, Sunny and Rish, as well as of my other family members, Subash, Monica, Riley and Aerin.

Professor Jagdish Handa
jagdish.handa@mcgill.ca

Part I

Introduction and heritage

1 Introduction

Monetary economics has both a microeconomics component and a macroeconomics one. The fundamental questions of monetary microeconomics concern the proper definition of money and its demand and supply, and those of monetary macroeconomics concern the formulation of monetary policy and its impact on the economy.

The financial assets that can serve the medium of the payments role of money have changed over time, as has the elasticity of substitution among monetary assets, so that the proper definition of money has also kept changing.

For short-run analysis, monetary economics is a central part of macroeconomics. The main paradigms of macroeconomics are the classical and Keynesian ones. The former paradigm studies the competitive economy at its full employment equilibrium, while the latter focuses on its deviations away from this equilibrium.

Key concepts introduced in this chapter

- Functions of money
- M1, M2, and broader definitions of money
- Financial intermediaries
- Creation of money by banks
- Classical paradigm for macroeconomics
- Walrasian general equilibrium model
- Neoclassical, traditional classical, modern classical and new classical models
- Keynesian paradigm for macroeconomics
- IS–LM analysis

Monetary economics is the economics of the money supply, prices and interest rates, and their repercussions on the economy. It focuses on the monetary and other financial markets, the determination of the interest rate, the extent to which these influence the behavior of the economic units and the implications of that influence in the macroeconomic context. It also studies the formulation of monetary policy, usually by the central bank or "the monetary authority," with respect to the supply of money and manipulation of interest rates, in terms both of what is actually done and what would be optimal.

In a monetary economy, virtually all exchanges of commodities among distinct economic agents are against money, rather than against labor, commodities or bonds, and virtually all loans are made in money and not in commodities, so that almost all market transactions in a

modern monetary economy involve money.[1] Therefore, few aspects of a monetary economy are totally divorced from the role of money and the efficiency of its provision and usage, and the scope of monetary economics is a very wide one.

Monetary economics has both a microeconomics and a macroeconomics part. In addition, the formulation of monetary policy and central bank behavior – or that of "the monetary authority," often a euphemism for the central banking system of the country[2] – is an extremely important topic which can be treated as a distinct one in its own right, or covered under the microeconomics or macroeconomics presentation of monetary economics.

Microeconomics part of monetary economics

The microeconomics part of monetary economics focuses on the study of the demand and supply of money and their equilibrium. No study of monetary economics can be even minimally adequate without a study of the behavior of those financial institutions whose behavior determines the money stock and its close substitutes, as well as determining the interest rates in the economy. The institutions supplying the main components of the money stock are the central bank and the commercial banks. The commercial banks are themselves part of the wider system of financial intermediaries, which determine the supply of some of the components of money as well as the substitutes for money, also known as near-monies.

The two major components of the microeconomics part of monetary economics are the demand for money, covered in Chapters 4 to 9, and the supply of money, covered in Chapter 10. The central bank and its formulation of monetary policy are covered in Chapters 11 and 12.

Macroeconomics part of monetary economics: money in the macroeconomy

The macroeconomics part of monetary economics is closely integrated into the standard short-run macroeconomic theory. The reason for such closeness is that monetary phenomena are pervasive in their influence on virtually all the major macroeconomic variables in the short-run. Among variables influenced by the shifts in the supply and demand for money are national output and employment, the rate of unemployment, exports and imports, exchange rates and the balance of payments. And among the most important questions in macroeconomic analysis are whether – to what extent and how – the changes in the money supply, prices and inflation, and interest rates affect the above variables, especially national output and employment. This part of monetary economics is presented in Chapters 13 to 20.

A departure from the traditional treatment of money in economic analysis is provided by the overlapping generations models of money. These have different implications for monetary policy and its impact on the economy than the standard short-run macroeconomic models.

1 Even an economy that starts out without money soon discovers its usefulness and creates it in some form or other. The classic article by Radford (1945) provides an illustration of the evolution of money from a prisoner-of-war camp in Germany during the Second World War.

2 In the United States and Canada, the control of monetary policy rests solely with the central bank, so that the central bank alone constitutes the "monetary authority". In the United Kingdom, control over the goals of monetary policy rests with the government while its implementation rests with the Bank of England (the central bank), so that the "monetary authority" in the UK is composed of the government in the exercise of its powers over monetary policy and the central bank.

While most textbooks on monetary economics exclude the overlapping generations models of money, they are an important new development in monetary economics. They are presented in Chapters 21 to 23.

The long-run analysis of monetary economics is less extensive and, while macroeconomic growth theory is sometimes extended to include money, the resulting monetary growth theory is only a small element of monetary economics. Monetary growth theory is covered in Chapter 24.

There are different approaches to the macroeconomics of monetary policy. These include the models of the classical paradigm (which encompass the Walrasian model, the classical and neoclassical models) and those of the Keynes's paradigm (which encompass Keynes's ideas, the Keynesian models and the new Keynesian models). We elucidate their differences at an introductory level towards the end of this chapter. Their detailed exposition is given in Chapters 13 to 17.

1.1 What is money and what does it do?

1.1.1 Functions of money

Money is not itself the name of a particular asset. Since the assets which function as money tend to change over time in any given country and among countries, it is best defined independently of the particular assets that may exist in the economy at any one time. At a theoretical level, money is defined in terms of the functions that it performs. The traditional specification of these functions is:

1 Medium of exchange/payments. This function was traditionally called the medium of exchange. In a modern context, in which transactions can be conducted with credit cards, it is better to refer to it as the medium of (final) payments.
2 Store of value, sometimes specified as a temporary store of value or temporary abode of purchasing power.
3 Standard of deferred payments.
4 Unit of account.

Of these functions, the medium of payments is the absolutely essential function of money. Any asset that does not directly perform this function – or cannot indirectly perform it through a quick and costless transfer into a medium of payments – cannot be designated as money. A developed economy usually has many assets which can perform such a role, though some do so better than others. The particular assets that perform this role vary over time, with currency being the only or main medium of payments early in the evolution of monetary economies. It is complemented by demand deposits with the arrival of the banking system and then by an increasing array of financial assets as other financial intermediaries become established.

1.1.2 Definitions of money

Historically, the definitions of money have measured the quantity of money in the economy as the sum of those items that serve as media of payments in the economy. However, at any time in a developed monetary economy, there may be other items that do not directly serve as a medium of payments but are readily convertible into the medium of payments at little cost

and trouble and can simultaneously be a store of value. Such items are close substitutes for the medium of payments itself. Consequently, there is a considerable measure of controversy and disagreement about whether to confine the definition of money to the narrow role of the medium of payments or to include in this definition those items that are close substitutes for the medium of payments.[3]

A theoretically oriented answer to this question would aim at a *pure* definition: money is that good which serves directly as a medium of payments. In financially developed economies, this role is performed by currency held by the public and the public's checkable deposits in financial institutions, mainly commercial banks, with their sum being assigned the symbol M1 and called the *narrow definition of money*. The checkable or demand deposits in question are ones against which withdrawals can be made by check or debit cards. Close substitutes to money thus defined as the medium of payments are referred to as *near-monies*.

An empirical answer to the definition of the money stock is much more eclectic than its theoretical counterpart. It could define money narrowly or broadly, depending upon what substitutes to the medium of payments are included or excluded. The broad definition that has won the widest acceptance among economists is known as (Milton) *Friedman's definition of money* or as *the broad definition of money*. It defines money as the sum of currency in the hands of the public plus all of the public's deposits in commercial banks. The latter include demand deposits as well as savings deposits in commercial banks. Friedman's definition of money is often symbolized as M2, with variants of M2 designated as M2+, M2++, or as M2A, M2B, etc. However, there are now in usage many still broader definitions, usually designated as M3, M4, etc.

A still broader definition of money than Friedman's definition is M2 plus deposits in near-banks – i.e. those financial institutions in which the deposits perform almost the same role for depositors as similar deposits in commercial banks. Examples of such institutions are savings and loan associations and mutual savings banks in the United States; credit unions, trust companies and mortgage loan companies in Canada; and building societies in the United Kingdom. The incorporation of such deposits into the measurement of money is designated by the symbols M3, M4, etc., by M2A, M2B, or by M2+, M2++, etc. However, the definitions of these symbols have not become standardized and remain country specific. Their specification, and the basis for choosing among them, are given briefly later in this chapter and discussed more fully in Chapter 7.

1.2 Money supply and money stock

Money is a good, which, just like other goods, is demanded and supplied by economic agents in the economy. There are a number of determinants of the demand and supply of money. The most important of the determinants of money demand are national income, the price level and interest rates, while that of money supply is the behavior of the central bank of the country which is given the power to control the money supply and bring about changes in it.

The *equilibrium amount* in the market for money specifies the *money stock*, as opposed to the *money supply*, which is a behavioral function specifying the amount that would be supplied at various interest rates and income levels. The equilibrium amount of money is the amount for which money demand and money supply are equal.

3 Goodhart (1984).

The money supply and the money stock are identical in the case where the money supply is exogenously determined, usually by the policies of the central bank. In such a case, it is independent of the interest rate and other economic variables, though it may influence them. Much of the monetary and macroeconomic reasoning of a theoretical nature assumes this case, so that the terms "money stock" and "money supply" are used synonymously. One has to judge from the context whether the two concepts are being used as distinct or as identical ones.

The control of the money supply rests with the monetary authorities. Their policy with respect to changes in the money supply is known as *monetary policy*.

1.3 Nominal versus the real value of money

The *nominal* value of money is in terms of money itself as the measuring unit. The *real* value of money is in terms of its purchasing power over commodities. Thus, the nominal value of a $1 note is 1 – and that of a $20 note is 20. The real value of money is the amount of goods and services one unit of money can buy and is the reciprocal of the price level of commodities traded in the economy. It equals 1/P where P is the average price level in the economy. The real value of money is what we usually mean when we use the term "the value of money."

1.4 Money and bond markets in monetary macroeconomics

The "money market" in monetary and macroeconomics is defined as the market in which the demand and supply of money interact, with equilibrium representing its clearance. However, the common English-language usage of this term refers to the market for short-term bonds, especially that of Treasury bills. To illustrate this common usage, this definition is embodied in the term "money market mutual funds," which are mutual funds with holdings of short-term bonds. It is important to note that our usage of the term "the money market" in this book will follow that of macroeconomics. To reiterate, we will mean by it the market for money, not the market for short-term bonds.

The usual custom in monetary and macroeconomics is to define "bonds" to cover all non-monetary financial assets, including loans and shares, so that the words "bonds," "credit" and "loans" are treated as synonymous. Given this usage, the "bond/credit/loan market" is defined as the market for all non-monetary financial assets. We will maintain this usage in this book except in Chapter 16, which creates a distinction between marketable bonds and non-marketable loans.

1.5 A brief history of the definition of money

The multiplicity of the functions performed by money does not aid in the task of unambiguously identifying particular assets with money and often poses severe problems for such identification, since different assets perform these functions to varying degrees. Problems with an empirical measure of money are not new, nor have they necessarily taken their most acute form only recently.

Early stages in the evolution from a barter economy to a monetary economy usually have one or more commodity monies. One form of these is currency in the form of coins made of a precious metal, with an exchange value which is, at least roughly, equal to the value of the metal in the coin. These coins were usually minted with the monarch's authority and were declared to be "legal tender," which obligated the seller or creditor to accept them in payment.

Legal tender was in certain circumstances supplemented as a means of payment by the promissory notes of trustworthy persons or institutions and, in the eighteenth and

nineteenth centuries, by bills of exchange[4] in Britain. However, they never became a generally accepted medium of payment. The emergence of private commercial banks[5] after the eighteenth century in Britain led to (private) note issues[6] by them and eventually also to orders of withdrawal – i.e. check – drawn upon these banks by those holding demand deposits with them. However, while the keeping of demand deposits with banks had become common among firms and richer individuals by the beginning of the twentieth century, the popularity of such deposits among ordinary persons came only in the twentieth century. With this popularity, demand deposits became a component of the medium of payments in the economy, with their amount eventually becoming larger than that of currency.

In Britain, in the mid-nineteenth century, economists and bankers faced the problem of whether to treat the demand liabilities of commercial banks, in addition to currency, as money or not. Commercial banking was still in its infancy and was confined to richer individuals and larger firms. While checks functioned as a medium for payments among these groups, most of the population did not use them. In such a context, there was considerable controversy on the proper definition of money and the appropriate monetary policies and regulations in mid-nineteenth century England. These disputes revolved around the emergence of bank demand deposits as a substitute, though yet quite imperfect, for currency and whether or not the former were a part of the money supply. Further evolution of demand deposits and of banks in the late nineteenth century and the first half of the twentieth century in Britain, Canada and the USA led to the relative security and common usage of demand deposits and established their close substitutability for currency. Consequently, the accepted definition of money by the second quarter of the twentieth century had become currency in the hands of the public plus demand deposits in commercial banks. During this period, saving deposits were not checkable and the banks holding them could insist on due notice being given prior to withdrawal personally by the depositor, so that they were not as liquid as demand deposits and were not taken to be money, defined as the medium of payments. Consequently, until the second half of the twentieth century, the standard definition of money was the narrow definition of money, denoted as M1.

Until the mid-twentieth century, demand deposits in most countries did not pay interest but savings deposits in commercial banks did do so, though subject to legal or customary ceilings on their interest rates. During the 1950s, changes in banking practices caused these savings deposits to increasingly become closer substitutes for demand deposits so that the major dispute of the 1950s on the definition of money was whether savings deposits should or should not be included in the definition of money. However, by the early 1960s, most economists had come to measure the supply of money by M2 – that is, as M1 plus savings

4 A bill of exchange is a promissory note issued by a buyer of commodities and promises to pay a specified sum of money to the seller on a specific future date. As such, they arise in the course of trade where the buyer does not pay for the goods immediately but is extended credit for the value of the goods for a short period, often three months. This delay allows the buyer time to sell the goods, so that the proceeds can be used to pay the original seller. In the nineteenth century, bills of exchange issued by reputable firms could be traded in the financial markets or discounted (i.e. sold at a discount to cover the interest) with banks. Some of them passed from hand to hand (i.e. were sold several times).

5 Many of the bankers were originally goldsmiths who maintained safety vaults and whose customers would deposit gold coins with them for security reasons. When a depositor needed to make a payment to someone, he could write a note/letter authorizing the recipient to withdraw a certain amount from the deposits of the payer with the goldsmith.

6 Private note issues were phased out in most Western countries by the early twentieth century and replaced by a monopoly granted to the central bank of the power to issue notes.

deposits in commercial banks – which does not include any types of deposits in other financial institutions. This mode of defining M2 is known as the Friedman definition (measure) of money, since Milton Friedman had been one of its main proponents in the 1950s and 1960s.

In the USA, during the 1960s, market interest rates on bonds and Treasury bills rose significantly above the ceilings set by the regulatory authorities on the interest rates that could be paid on saving deposits in commercial banks. Competition in the unregulated sphere led to changes in the characteristics of existing near-monies in non-bank financial intermediaries which made them closer to demand deposits and also led to the creation of a range of other assets in the unregulated sphere. Such liabilities of non-financial intermediaries were substitutes – some closer than others but mostly still quite imperfect ones – for currency and demand deposits. Their increasing closeness raised the same sort of controversy that had existed during the nineteenth century about the role of demand deposits and in the 1950s occurred about savings deposits in commercial banks. Similar evolution and controversies occurred in Canada and the UK. The critical question in these controversies was – and still is – how close does an asset have to be to M1, the primary medium of payments, to be included in the measure of money.

Evolution of money and near-monies since 1945

To summarize the developments on the definition of money in the period since 1945, this period opened with the widely accepted definition of money as being currency in the hands of the public plus demand deposits in commercial banks (M1). This definition emphasized the medium of payments role of money. Demand deposits were regulated in several respects, interest could not be legally – or was not customarily – paid on them, and certain amounts of reserves had to be legally – or were customarily – maintained against them in the banks. Against this background, a variety of developments led to the widespread creation and acceptance of new substitutes for demand deposits and the increasing closeness of savings deposits to demand deposits. In Canada, this evolution increased the liquidity of savings deposits with the chartered banks, which dominated this end of the financial sector, with also some increase in the liquidity of the liabilities of such non-monetary financial institutions as trust companies, credit associations[7], and mortgage and loan associations. In the United States, until the 1970s, the changes increased the liquidity primarily of time deposits in the commercial banks, and to some extent of deposits in mutual savings banks, and shares in savings and loan associations. In the United Kingdom, the increase in liquidity occurred for interest-bearing deposits in retail banks and building societies. Given this evolution in the 1960s and 1970s, a variety of studies established these assets to be fairly close – but not perfect – substitutes for demand deposits.

This evolution of close substitutes for M1 led in the 1950s to a renewal of controversy, almost dormant in the first half of this century, on the proper definition of money. In particular, in the third quarter of the twentieth century, there was rapid growth of savings deposits in commercial banks and in non-bank financial intermediaries, with their liabilities becoming increasingly closer substitutes for demand deposits, without their becoming direct media of payments. This led to the acceptance of M2 as the appropriate definition of money, though not without some disputes. In the fourth quarter, as mentioned above, there have been numerous innovations that have made many liabilities of financial intermediaries increasingly

7 An example of these is *caisses populaires* in Quebec, Canada.

indistinguishable from demand deposits. This has led to the adoption or at least espousal of still wider definitions under the symbols M3, M4, etc.

Financial innovations

Financial innovation has been extremely rapid since the 1960s. It has included technical changes in the servicing of various kinds of deposits, such as the introduction of automatic teller machines, telephone banking, on-line banking through the use of computers, etc. It has also included the creation of new assets such as Money Market Mutual Funds, etc., which are often sold by banks and can be easily converted into cash. There has also been the spread first of credit cards, then of debit or bank cards, followed still more recently by the attempts to create and market "electronic money" cards – sometimes also known as electronic purses or smart cards. Further, competition among the different types of financial intermediaries in the provision of liabilities that are close to demand deposits or are readily convertible into the latter, increasingly by telephone and online banking, has increased considerably in recent decades. Many of these innovations have further blurred the distinction between demand and savings deposits to the point of its being only in name rather than in effect, and also blurred the distinction between banks and some of the other types of financial intermediaries as providers of liquid liabilities. This process of innovation, and the evolution of financial institutions into an overlapping pattern in the provision of financial services, are still continuing.

Credit cards allow a payer to pay for a purchase while simultaneously acquiring a debt owed to the credit card company. Because of the latter, most economists choose not to include credit card usage or their authorized limits in the definition of money. Nor are credit cards near-monies. However, their usage reduces the need for the purchaser to hold money and reduces the demand for money.

Debit cards are used to pay for purchases by an electronic transfer from the buyer's bank account, often a demand deposit account with a bank. They replace the need to make payments in currency or by issuing a check. Therefore, they reduce currency holdings. They also reduce payments by checks. However, they do not obviate the need to hold sufficient balances in the bank account on which the debit is made. They are expected to have a very limited impact on the holding of deposits, which could increase or decrease.

Electronic transfers are on-line transfers made over the Internet. They reduce the need to use checks for making payments. However, electronic transfers may not affect deposits in banks, or do so marginally due to better money-management practices afforded by on-line banking.

Smart cards embody a certain cash value and can be used to make payments at the point of purchase. Given the increasing prevalence of online banking and debit cards, smart cards are likely to be mainly used for small payments, as in the case of telephone cards, library photo-copying cards, etc. Smart cards reduce the need to hold currency and reduce its demand.

Therefore, financial innovations in the form of debit and smart cards reduce currency holdings rather than demand deposits. Financial innovations in the form of online transfers facilitate the investment of spare balances, which at one time may have been held in savings deposits, in higher-interest money market funds, etc., thereby reducing the demand for savings deposits.

In recent decades, the reduction in brokerage fees for transfers between money and non-monetary financial assets (bonds and stocks) and the Internet revolution in electronic banking have meant a reduction in the demand for money. Part of this is due to a reduction in the

demand for precautionary balances held against unexpected consumption expenditures. This reduction has taken place because individuals can more easily and at lower cost accommodate unexpected expenditure needs by switching out of other assets into money.

Theoretical and econometric developments on the definition of money

Keynes in 1936 had introduced the speculative demand for money as a major motive for holding money and Milton Friedman in 1956 had reinterpreted the quantity theory of money to stress the role of money as a temporary abode of purchasing power, similar to a durable consumer good or a capital good. This analysis is presented in Chapter 2. Numerous theoretical and empirical studies in the 1950s and 1960s pointed out the development of close substitutes for money as a feature of the financial evolution of economies. By the 1960s, these developments led to a realignment of the functional definition of money to stress its store of value aspect, in this case as an asset relative to other assets, rather than medium of payments aspect. The result of this shift in focus was to further stress the closeness of substitution between the liabilities of banks and those of other financial intermediaries.

Such shifts in the definition of money were supported both by shifts in the analysis of the demand for money, suited to the stress on the store-of-value function, and by a large number of empirical studies. However, in the presence of a variety of assets performing the functions of money to varying degrees, purely theoretical analysis did not prove to be a clear guide to the empirical definition or measurement of money. As a result, research on measuring the money stock for empirical and policy purposes took a variety of routes after the 1960s. Several broad routes may be distinguished in this empirical work. Two of these were:

1 One of the routes was to measure money as the sum of M1 and those assets that are close substitutes for demand deposits. Closeness of substitution was determined on the basis of the *price and cross-price elasticities* in the money-demand functions or of the *elasticities of substitution* between M1 and various non-money assets. Such studies, discussed in Chapter 7, generally reported relatively high degrees of substitution among M1, savings deposits in commercial banks, and deposits in near-bank financial intermediaries and therefore supported a definition of money that is broader than M1 and in many studies even broader than M2.
2 The second major mode of defining money was to examine its appropriateness in a macroeconomic framework. This analysis is presented in Chapter 9. In this approach, the definition of money was specified as that which would "best" explain or predict the course of nominal national income and of other relevant macroeconomic variables over time. But there proved to be little agreement on what these other relevant variables should be. The quantity theory tradition (in the work of Milton Friedman, most of his associates and many other economists) took nominal national income as the only relevant variable. For the 1950s and 1960s, this approach found that the "best" definition of money, as shown by examining the correlation coefficients between various definitions of money and nominal national income, was currency in the hands of the public plus deposits (including time) in the commercial banks. This was the Friedman definition of money and was widely used in the 1960s. However, it should be obvious that the appropriate definition of money under Friedman's procedure could vary between periods and countries, as it did in the 1970s and 1980s.

Further, in the disputes on this issue in the 1960s, many researchers in the Keynesian tradition took the appropriate macroeconomic variables related to money as being nominal national income and an interest rate, and defined money much more broadly than M2 to include deposits in several types of non-bank financial intermediaries and various types of Treasury bills and government bonds.

Up to the 1970s, empirical work along the above lines brought out an array of results, conflicting in detail though often in agreement that M2 or a still wider definition of money performs better in explaining the relevant macroeconomic variables than money narrowly defined. This consensus vanished in the 1970s and 1980s in the face of increasing empirical evidence that none of the simple-sum aggregates of money – whether M1, M2 or a still broader one – had a stable relationship with nominal national income. Research on the 1970s and 1980s data showed that (a) the demand functions for the various simple-sum monetary aggregates were unstable, and (b) they did not possess a stable relationship with nominal income.

The above findings for the simple sum aggregates prompted the espousal of several new functional forms for the definition of money. Among these are the Divisia aggregates. The construction of and comparison between different monetary aggregates is the subject of Chapter 7. The search for stability of the money-demand function also led to refinement of econometric techniques, resulting in cointegration analysis and error-correction modeling of non-stationary time series data, and the derivation of separate long-run and short-run demand functions for money. These issues are further examined in Chapter 9.

Further, the continuing empirical instability of the demand functions for M2 and still broader definitions of money since the 1980s led to an increased preference for some form of M1 over broader aggregates for policy formulation and estimation, thereby reversing the shift towards M2 and other broad monetary aggregates which had occurred in the 1950s and 1960s. Further, the empirical instability of money-demand functions led to a marked decrease after the 1980s in both analytical and empirical studies on the definition of money.

In addition, after the 1980s, at the monetary policy and macroeconomic level, many central banks and researchers have chosen to focus on the interest rate as the appropriate monetary policy instrument – thereby relegating money supply and demand to the sidelines of macroeconomic reasoning. The discussion of this shift and its implications for macroeconomic modeling and policy analysis is to be found in Chapters 13 to 15.

1.6 Practical definitions of money and related concepts

We have already referred to several definitions of money. These definitions are fairly, though not completely, standardized across countries for M1 and M2 but tend to differ for broader designations. The generic definitions of these monetary variables can be taken to be as follows:

- M1 = Currency in the hands of the public + checkable deposits in commercial banks;
- M2 = M1 + savings deposits in commercial banks.

These generic definitions are modified to suit the context of different countries and their central banks. Further, in general, with increases in the substitutability of different monetary assets, the definitions of each of the aggregates have broadened over time. Often, the variations in the definition of M1 are accommodated by using terms such as M1, M1+, M1++, etc.

As an illustration of the variations in the various measures of money in practice in different countries, the following gives the current definitions of the major monetary aggregates in the USA:

- M1 = currency in circulation among the public (i.e. excluding the Fed, the US Treasury and commercial banks) + demand deposits in commercial banks[8] (excluding interbank and US government deposits and those of foreign banks) + other checkable deposits including negotiable orders of withdrawal (NOW) + credit union (such as Savings and Loan Associations) share draft accounts + demand deposits at thrift institutions (such as Mutual Savings Banks) – cash items in the process of collection and Federal Reserve float;
- M2 = M1 + savings deposits, including money market deposit accounts + small time deposits under $100,000 + balances in retail money market mutual funds;
- M3 = M2 + time deposits over $100,000 + Eurodollars held by US residents at foreign branches of US banks and at all banks in the UK and Canada + money market mutual funds held by institutions.

Note that M1, M2 and M3 exclude amounts held by US commercial banks, the US government, money market funds, foreign banks and official institutions.

The above detailed descriptions for the United States of M1 and M2 are more complex than our usual modes of defining them. However, our usual definitions are reasonable proxies. Under our proxy definitions, M1 is defined as currency in the hands of the public plus checkable deposits in deposit-taking financial institutions. M2 is defined as M1 plus (small or retail) time and savings deposits in these institutions.

For Canada, the monetary aggregates are measured as:

- M1 = currency in the hands of the public and demand deposits in chartered banks[9];
- M1+ = M1 + personal checkable deposits + non-personal checkable notice deposits at chartered banks, mortgage loan companies and credit unions;
- M2 = M1 plus personal savings deposits and non-personal notice deposits at chartered banks;
- M2+ = M2 plus deposits at trust and mortgage loan companies and credit unions (including *caisses populaires*[10]);
- "Adjusted M2+" = M2+ plus Canada Savings Bonds and mutual funds at financial institutions;
- M3 = M2 plus non-personal fixed-term deposits at chartered banks and foreign currency deposits of residents booked in Canada.

For the United Kingdom, the definitions of the symbols in common usage are:

- M1 = currency plus current account (checking) sterling deposits in retail banks and building societies, held by "UK residents"[11];

8 Our usage of the term "commercial banks" refers to "depository institutions" in the USA.
9 The chartered banks in Canada correspond to the commercial banks in our discussions.
10 These are essentially credit unions in Quebec.
11 "UK residents" is meant to exclude the public sector and the financial institutions.

- M2 = currency plus sterling deposits in retail banks[12] and building societies, held by UK residents;
- M4 = currency plus sterling deposits at the central bank, other banks and building societies held by UK residents.

Note that the definitions of the monetary aggregates beyond M2, e.g. M3 and M4, differ more radically among countries than those of M1 and M2. For M3 and M4, the only common denominator is that they are broader than M2 and include, besides M2, other highly liquid assets held at financial institutions. The reliance on these specific wider aggregates usually reflects the peculiarities of the country's financial structure.

Note also that currency holdings and M1 are becoming increasingly smaller proportions of M2 and wider aggregates. In the USA, at the end of 1995, the amount of currency in the economy was \$379b[13], M1 was \$1150b, M2 was \$3680b and M3 was \$4954b. The ratio of M1 to M2 was only 31 percent and to M3 was 23 percent. For Canada in 1995, the currency in the economy was \$26.8b, M1 was \$62.7b, while M2+ was \$618.4b The ratio of M1 to M2+ was only 10 percent. For the UK in 1995, currency holdings were £20.8b, M2 was £439.4b, and M4 was £682.5b.[14]

1.6.1 Monetary base and the monetary base multiplier

The money supply is related to the monetary base – sometimes called the *reserve base* – by the monetary base multiplier. Since this multiplier is greater than one, the monetary base is also known as *high-powered money*. We will use the symbol M0 for it. Its generic definition is:

- M0 = Currency in the hands of the non-bank public plus currency held by the commercial banks + reserves held by the commercial banks with the central bank.

The central bank can control the monetary base through open market operations and other measures, for which see Chapter 11. For any given definition of money, the "monetary base multiplier" is defined as $\partial M/\partial M0$. If the value of this multiplier is constant or a function of a small set of variables, the central bank may be able to control the money supply by changing the monetary base. However, our remarks in this chapter on the instability of the money-demand function in recent decades imply that this multiplier is definitely not a constant or even a stable function of a small set of variables because of extensive financial innovation, so that the central bank's control over the monetary base has not ensured a similar degree of control over the money supply.

The monetary base (to money) multiplier needs to be distinguished from the "money (to nominal income) multiplier," which is defined as $\partial Y/\partial M$, where Y is nominal national income. Since $Y \equiv MV$, where V is the velocity of money (see Chapter 2), the money multiplier equals V. This multiplier is normally not a constant but, at the minimum, is a function of several variables, including the interest rate. This function may, or may not be, also unstable.

12 Since 1993, these deposits include both non-interest-bearing and interest-bearing deposits.

13 b stands for billion, defined as 1000 million.

14 These figures are taken from *Statistics on Payment Systems in the Group of Ten Countries*, published by the Bank for International Settlements, various years.

Therefore, the central bank's control over the monetary base need not ensure a high degree of control over nominal income because of the instability of the monetary base multiplier or the instability of the velocity of circulation of money, or both.

1.7 Interest rates versus money supply as the operating target of monetary policy

Central banks may exercise control over the economy's interest rates, in addition to the money supply or, in its place, as their monetary policy instrument. Our concern at this point is really with the instrument that is used as the primary one – that is, set exogenously by the central bank. If the money supply is used as the primary instrument, the economy's interest rates will change in response to the central bank's changes in the money supply, and will be endogenous. If the interest rates are used as the primary monetary policy instrument, the economy's money demand will change in response to the changes in interest rates. In this case, if the money market is to maintain equilibrium, the central bank has to accommodate the change in money demand by appropriate changes in the money supply, so that the money supply will become endogenous. While the choice between the money supply and the interest rates can be trivial under certainty and a stable money-demand function, it is not likely to be trivial under uncertainty and an unstable money-demand function, so that central banks are forced to make choices between the two alternatives.

This issue has been brought to the forefront of the debate on the appropriate macroeconomic analysis by the adoption by central banks in several developed economies of the policy of using the interest rate as the primary monetary policy instrument – and the abandonment of this role for the money supply. Several new Keynesian models of the last two decades now incorporate such an assumption. This issue will be discussed in proper detail in the presentation of these models later in this chapter and in Chapters 13 and 15.

1.8 Financial intermediaries and the creation of financial assets

Asset transmutation by financial intermediaries

Financial intermediaries are institutions that *intermediate* in the financial process between ultimate borrowers and ultimate lenders in the economy. The *ultimate borrowers* include (a) consumers who need to borrow to finance part or all of their consumption, (b) firms that borrow to invest in physical capital, and (c) the government when it borrows to finance its deficits. The *ultimate lenders* are the economic units that save part of their current income by spending less than their current income on their purchases of commodities and want to lend some or all of these savings to others for some duration. Householders form the major bulk of the ultimate lenders, saving part of their current income. Some of the firms engaged in production also do not spend all of their sales revenue on immediate purchases of inputs or distribute them to shareholders as distributed profits but save part of them (i.e. keeping some profits as retained earnings). They are sometimes willing to lend part of these retained earnings to others. The government does the same on a net basis when it runs a surplus.

Financial intermediaries borrow from the ultimate lenders or from other intermediaries by issuing their own liabilities in exchange and re-lend to others by accepting the latter's liabilities. In the modern economy, only a small proportion of the savings is directly transferred from the savers to the ultimate borrowers. Most of the savings are directed by the savers to financial intermediaries such as banks, mutual funds, pension funds, insurance

companies, etc., which re-channel the funds thus obtained to firms and the government, either directly by buying their shares and bonds or indirectly through other financial intermediaries such as investment banks.

The basic reason for this intermediation is the differences in the preferences of the savers for asset characteristics, such as liquidity and security, and those attaching to the instruments issued by the firms and the government. Consequently, there is in general a considerable difference in the characteristics of the liabilities sold to the savers by a financial intermediary and those of the assets bought by it, resulting in what is sometimes called the *asset–transmutation process*.

Banks are financial intermediaries that borrow from the public by inviting demand and time deposits or issuing their own securities and hold the liabilities issued by others. Their existence is a superb example of asset transmutation through financial intermediation. The main liabilities of banks are deposits which are virtually riskless to the depositors since they are payable on demand or after a short specified notice. In short, they are highly liquid. By contrast, the assets held by the banks are government securities, loans to the public, etc., possessing some risk of loss and, as with loans, a limited degree of marketability or encashment at short notice. Therefore, the assets issued by the banks are much more liquid than the assets held by them. Conversely, the return paid by the banks on the former is less than the return that the banks earn on the latter.

Multiple creation of financial assets

All financial assets are "created" and have no intrinsic physical existence, but are the liabilities of some economic unit or other. They may be examined in terms of their characteristics, especially in terms of their yield or expected yield, risk of loss, marketability, maturity and so on. Anyone purchasing a financial asset may be thought of as purchasing a particular set of characteristics, such as risk and marketability etc., in exchange for a specified expected yield on the asset. Financial intermediaries cater to this demand through the creation of assets with differing combinations of characteristics. For many pairs of assets, it is feasible for some intermediary to create a third asset that offers a mix of the characteristics of the original two assets, so that the multiplicity of differentiated assets is a common outcome of unregulated financial intermediation.

Financial intermediaries typically issue assets with more desirable characteristics for lenders than do the ultimate borrowers, persuading the latter to hold the liabilities of the intermediaries. In turn, the intermediaries use the funds obtained from the sale of their liabilities to purchase the liabilities of other borrowers which pay a higher expected net yield, thus covering the expenses of intermediation and making a profit in the process.

Financial intermediaries permeating an unregulated economy lead to a multiplicative creation of their liabilities. To illustrate, consider an economy in which everyone is willing to hold the asset A issued by a given intermediary.[15] Now assume that an ultimate lender saves \$100 and exchanges it for the asset A. The intermediary transfers (lends) the \$100 to another individual B, who transfers them in some way, such as through consumption or investment expenditures, to a third individual C. The last individual again exchanges the \$100 of funds for the assets issued by the intermediary. Suppose these are the only transactions that take place in a given period and there are no leakages at any point. Then, the intermediary, for an

15 For example, deposits in a bank.

initial $100 lent to it, has created $100 of its liabilities. The amount created over *n* periods will be $100*n* and will approach infinity over time. The implication of this example is clear: the multiplicative creation of the liabilities of financial intermediaries is inherent in an economy in which these liabilities are widely held. The extent of this creation is limited by the leakages out of the recycling process. Thus, if individual C had only deposited $50 of resources with the intermediary and retained the remainder in his storage, the recycling process would have had a leakage of $50 (or 50 percent). The total assets created by the intermediary in the period would be worth only $50 and only $100 over time.

Banks conform to the above pattern. The funds they receive are deposits of currency and are part of their reserve base. They lend these out, after keeping some of this currency to meet their own demand for reserves, part of which they are in some countries legally required to keep. The public receiving the funds may, after some transfers within its own members or even without any transfers, redeposit the funds in the banks. It may also retain some currency to meet its own demand for it. The remainder returns to the banks and starts the next cycle of the asset–creation process. The leakages in the form of the currency demand of the public and of the banks against their deposits prevent an infinite creation of deposits over time but nevertheless lead to some multiple expansion of the banks' liabilities, unless the leakages were 100 percent in the first cycle.

Since financial assets are created, it is natural to expect that in an unregulated system a variety of financial assets differing only slightly in their characteristics and with varying degrees of closeness of substitution will come into existence. Further, any regulation of existing assets tends to increase the profitability of unregulated potential substitutes and usually leads to their creation. These tendencies towards multiplicity of financial assets introduce severe problems in defining money and in its regulation. Further, the financial development of a country is usually marked by the increasing richness of the financial assets available in it and the increasing closeness of near-money assets with the asset that serves as the medium of payments. Consequently, *questions on the proper definition of money never seem to die out*, thereby posing a continual challenge to monetary economists to appropriately redefine money.

Distinctive role of banks as financial intermediaries

Banks are not the only financial intermediaries in the economy. But they are the most widespread and their liabilities are so widely demanded that the multiple creation of their liabilities is both the greatest and the most widely recognized. Banks, accepting demand and time deposits, differ from other financial intermediaries in that their liabilities are readily acceptable and are liquid since demand deposits are a medium of payments and hence a form of money. Further, another of their liabilities, time deposits, is a very close substitute for currency and demand deposits. By comparison, the liabilities of non-bank financial intermediaries are not directly a medium of payments, nor are they perfect substitutes for it. This special role of the liabilities of banks in the economy makes the banks a rather distinctive type of financial intermediary and makes a study of their behavior and reaction to monetary policy especially important.

Fragility of the financial system

The financial system is said to be fragile in the sense that it is prone to crisis. The reasons for this include banks' reliance on fractional reserves and asset transmutation. Since they

hold only a small part of their liabilities, mainly deposits, in reserves (currency holdings and deposits with the central bank), they are not able to refund these deposits if the depositors wish to suddenly and simultaneously try to withdraw a considerable fraction of deposits. Such a withdrawal from a bank is known as a "run" on the bank and its most visible manifestation is long lines of depositors waiting to enter the bank to make their withdrawals. Asset transmutation means that banks' liabilities have a much shorter maturity than their assets. In the case of a run on the bank, if the bank tries to sell its assets at short notice it is likely to incur a loss over the amount that it would get if it held the assets to maturity or could sell them at a more opportune time. Also, note that a considerable part of the assets held by banks are in the form of non-marketable loans, which are difficult to convert into cash at short notice.

Therefore, the fractional reserve system, with its asset transmutation, rests on trust by the depositors in the continuing liquidity and solvency of the bank in question. The emergence of less than absolute trust in the bank's ability to honor withdrawals from it, even if this is due to an unjustified rumor or just contagion spreading from other financial institutions, can be enough to trigger a run on it, as well as a refusal by other financial institutions to come to its aid and lend to it. This can soon result in closure of the bank. Some protection against such an eventuality is provided by insurance of its deposits, often by a public agency, so that depositors do not need to worry about the safety of their deposits, and by the central bank's doctrine of "lender of last resort." Under the latter, even if no private lender will lend to the bank, the central bank will do so. These issues are covered in Chapter 11 on central banking.

1.9 Different modes of analysis of the economy

Since the money market is only one of the markets in the economy, monetary economics is closely intertwined with the analysis of the other markets in the economy. This unified analysis of money and all other markets in the economy can be conducted in one of two ways:

(I) *A microeconomic analysis* of the market for each of the goods in the economy. While there can be different types of such models, many of them are made analytically tractable at the level of the economy by imposing the assumptions of perfect markets (perfect competition and instant market clearance), absence of market imperfections such as frictions and transactions costs, etc., on the analysis of each market. Other types of microeconomic models discard one or more of these assumptions for selected markets.

Microeconomic models of the economy assuming perfect competition are called *Walrasian models*. They are difficult to manage unless the assumption of equilibrium – that is, demand equal to supply – in all markets is imposed on them. The (subsidiary) group of Walrasian models that does so provides microeconomic models of the economy known as the *Walrasian general equilibrium models*. Given the assumptions of perfect competition, absence of frictions, transactions costs and uncertainty, as well as general equilibrium in all markets, including the labor market, such a model implies that, in the general equilibrium state, money is neutral, that is, changes in the money supply do not alter the values of the real variables, including employment and the output of commodities. This equilibrium is usually called the "long-run state" of the model. However:

(i) Money is is not neutral in most specifications, in cases where they are spelled out, of the disequilibrium or the short-run equilibrium states of the Walrasian models.

(ii) Money is not neutral even in the equilibrium of those models that dispense with one or more of the assumptions listed above. Such states are often labeled as "short-run" equilibrium ones. For instance, it is not neutral in short-run equilibrium if there is uncertainty and errors in expectations, or if the markets are not perfect, or if frictions and transactions costs exist.

(II) *A macroeconomic analysis*, where the goods are classified into a small number of categories and the analysis is performed at this composite level. Although many different ways of categorizing goods are possible, the one generally used in short-run macroeconomics is that of classifying goods for the closed-economy analysis into the four categories of commodities, money, bonds (non-monetary financial assets), and labor, and that of open-economy analysis into the above four goods and foreign exchange.

The relationship between the microeconomic and the macroeconomic varieties of models can be either:

(A) (II) is merely a compact form of (I). In this case, the assumptions and implications of macroeconomic analysis must be consistent with the microeconomic analysis of markets. This approach seeks to set the foundations of macroeconomics in microeconomic theory. Note that doing so will only embody whatever features the underlying microeconomic model possesses. Therefore, if the underlying model possesses nominal wage and/or price rigidities or allows the absence of instantaneous market clearance, so will the derivative macroeconomic model.[16] If the model assumes the absence of nominal wage and price rigidities and instant market clearance, so will the derivative model.[17]

(B) (II) is different from and possibly more insightful than just being a compact form of (I). In this case, in addition to the assumptions on the individualistic behavior of economic units, macroeconomic models can incorporate assumptions that deal with group behavior,[18] as well as interactions among markets[19] and groups[20] that are not visible in microeconomic analysis. If these are relevant, macroeconomics provides the guide for the specification of the appropriate microeconomic analysis, so that macroeconomics serves as the foundation for the relevant microeconomics. Further, models of type (II) are often more tractable for studying the properties of the economy in disequilibrium or if there exist departures from perfect competition.

Each of the above types of analysis has its advantages and disadvantages. The advantage of (A) is that it roots macroeconomic behavior in the microeconomic analysis of the household and firm, which provides a check on the rationality of the assumed behavior of households and firms. However, there are two major disadvantages to using (A). One of these is that it usually extends the assumption of continuous equilibrium to all markets, thereby usually assuming

16 New Keynesian macroeconomic models of the last two decades are usually of this type.

17 Modern versions of the classical school tend to be of this type.

18 For example, a "herd instinct" (such as contagion, panic and euphoria in stock markets) may be important when studying the behavior of groups but not when studying the behavior of any one economic agent. While the existence of herds is intuitively obvious, its formal modeling is still at an early stage.

19 For example, spillovers between markets can be quite important between the labor market and the commodity market – such as the fact that unemployed workers reduce their consumption of commodities – while they tend to be ignored in the microeconomic analysis of individual economic agents and markets.

20 E.g. labor unions and firms' cartels.

that all markets are simultaneously, and always, in equilibrium. While such an assumption seems quite sensible and may be relatively innocuous at the level of one market, it is often not a sufficiently valid assumption for the whole economy. In particular, the assumption of simultaneous and instantaneous equilibrium prevents the study of the elements of the pathology of the system,[21] i.e. when some part of it breaks down, so that the overall system does not possess general equilibrium, possibly not even the ability to return to it soon in real time.[22] The other major disadvantage, as mentioned above, of purely microeconomic analysis is that it tends to ignore behavior that is applicable only in the mass or in groups but not to individual economic units studied in isolation.[23]

The Walrasian general equilibrium system provides the *benchmark* of the well-functioning, healthy economy. It is extremely useful in this respect and remains central to the study of macroeconomics. Among the major components of this system are: a complete set of markets for all possible goods, utility maximization by consumers and workers and profit maximization by firms,[24] perfectly competitive and perfectly efficient markets,[25] certainty or the absence of errors in expectations, absence of barriers to the attainment of equilibrium, absence of lags and "false trading,"[26] and the availability of a mechanism for instantly reaching the general equilibrium for the economy.[27] This is indeed a tall set of assumptions and economists use them mainly for deriving implication for the analytical long-run state of the economy. They approach the short-run as a minor deviation from the long-run model. Macroeconomic models of this type are the "classical" group of models. They belong to category (A) above.

The Walrasian general equilibrium system, incorporating perfect competition and perfect efficiency, by its very nature, does not provide an appropriate platform for studying the pathology of the economy when it is not functioning well in whole or in some of its parts. Their main rival is the Keynesian group of models, which focus on the pathology of the economy to specify the short-run analysis of the economy, with the long-run model becoming a variation on short-run modeling. Keynesian models belong to category (B) above.

1.10 The classical paradigm: the classical group of macroeconomic models

The classical group of models is consistent with the Walrasian general equilibrium framework and assumes that the market establishes the wages and prices for each of the goods at that

21 A comparison of the economy with the human body can illustrate this point. The human body does not always stay healthy. Further, if it gets sick, it may be able to recover back to good health but may not do so soon. Hence, modeling (studying) only the properties of the healthy body may provide poor or disastrous recommendations of what treatment to administer when it does become ill.

22 Here, the chronological time taken to reach the general equilibrium state is of the essence, so that the properties of long-run equilibrium are of little consequence unless the chronological time taken to reach the long-run is also specified.

23 Note that neither of these disadvantages resorts to an appeal to irrational economic behavior. Additionally, if behavior is, in fact, non-rational, then a model which excludes such behavior would not capture reality.

24 Under uncertainty, these would become expected utility maximization by both households and firms.

25 Markets are perfectly "efficient" if they instantaneously restore equilibrium following a shock to demand and/or supply. Note that this is a different assumption from that of perfect competition, which is defined as the state in which no buyer or seller can influence the market price.

26 That is, trading at prices and quantities other than those that will occur in equilibrium.

27 These are often presented under the rubric of the Arrow–Debreu model, which is a rigorous statement of the Walrasian general equilibrium model.

level at which its notional demand and supply are equal (i.e. at which its market "clears"). Since one of the markets is labor, its clearance implies that every worker who wishes to supply labor at the existing wage will have a job and each firm will be able to employ all the workers that it wants to at the existing wage. This state, in the context of the long-run analysis, is known as "full employment," so that a hallmark of the classical models is that, in long-run equilibrium, they imply full employment.[28] However, in view of their emphasis on labor market clearance, this implication of equilibrium is often turned around and stated as if it was an assumption, which is not strictly correct.[29]

While there is no consensus on the division of the classical group of models into individual models, we adopt the following taxonomy for this book.

I. Traditional classical ideas

"The traditional classical approach (or ideas)" is being proposed in this book as the name for the somewhat disparate ideas on the macrostructure of the economy from the middle of the eighteenth century to the publication of Keynes's *The General Theory* in 1936. To quite a considerable extent, these ideas were diffuse, varied among authors and changed over time. In any case, there was no single compact version of the overall exposition, though the profession, following Keynes, now treats them as if there was a compact model. We will call this compact statement of the traditional classical ideas the traditional classical model. It was never stated as a compact model even during its heyday during the nineteenth and early twentieth centuries, but its ideas permeate the classical paradigm.

The two components of the traditional classical model directly relevant to monetary economics were the *quantity theory* for the determination of prices (see Chapter 2) and the *loanable funds theory* for the determination of interest rates (see Chapter 19). Its *theory of employment* was the analysis of the labor market and incorporated the assumption of equilibrium, which state represents full employment, so that the traditional classical set of ideas did not possess a theory of unemployment or of variations in aggregate employment other than those of variations in their long-run levels. Hence, it did not possess a theory of the deviations in unemployment and output from their full-employment levels. However, another component of the traditional classical ideas was its *business cycle explanations*, which allowed for fluctuations in economic activity in the economy's response to real or monetary shocks, so that such explanations implicitly did envisage deviations from full employment.

The traditional classical approach lacked the integration of its microeconomic-based theory of employment and output with its business cycle explanations, as well as of their mix with the quantity theory and the loanable funds theory. To sum up, while this approach had many of the components of macroeconomics, it lacked an integrated macroeconomic framework. It also lacked an explicit treatment of the aggregate demand for commodities, now encompassed in the IS relationship, which is an essential building block of current macroeconomics.

28 In the short-run of the classical models, with errors in expectations, market clearance can imply a level of employment different from the full-employment (long-run equilibrium) level.

29 The difference between an assumption of full employment and one that is an implication of the equilibrium state is that the former rules out the studies of the properties of the system when it is in disequilibrium; the latter does not necessarily do so. In addition, the former can rule out a distinction between short-run and long-run equilibrium.

THEORY OF AGGREGATE DEMAND IN THE TRADITIONAL CLASSICAL APPROACH AND SAY'S LAW

There was also no explicit macroeconomic theory of the commodity market in the traditional classical approach since it did not incorporate a theory for the determination of the aggregate demand for commodities.[30] Instead, this approach studied each commodity market separately in microeconomic terms, that is, in terms of its demand and supply analysis. In place of a theory of aggregate demand for commodities as a whole, the traditional classical approach explicitly, but more often implicitly, settled for Say's law (see Chapter 18), which stated that, in the aggregate, the supply of commodities creates (i.e. always generates) its own demand, so that a separate theory of aggregate demand was not needed or specified.

Say's law was pervasive in the analyses offered by many economists throughout the classical period: among others, it was espoused by Adam Smith in the eighteenth century, David Ricardo in the early nineteenth century, John Stuart Mill in the mid-nineteenth century and Alfred Marshall in the late nineteenth century. However, Say's law is not valid for a monetary economy, which possesses commodities, money and bonds, for several reasons. One of these is that, in a monetary economy, all sellers of commodities are not automatically buyers of commodities to the same extent, since a part of the income of sellers is usually saved, which can be put by them into money or bonds (which include savings deposits in banks) rather than being automatically converted into spending on commodities.

Note that modern theories of aggregate demand do not embody Say's law, so that it is no longer a part of modern macroeconomics.

II. Neoclassical model

The "neoclassical model" is the name given to the restatement of the traditional classical ideas rebottled and re-flavored in the post-*General Theory* period in a new compact framework. The new bottle was the *IS–LM framework* of analysis; the re-flavoring included the elucidation of some of the nuances of the traditional classical ideas, such as the wealth/Pigou and real balance effects (see Chapter 3) on commodity demand, as well as the addition of new elements such as the speculative demand for money (see Chapter 5) and the explicit analysis of the commodity market at the macroeconomic level. Further, certain elements of the traditional ideas such as the quantity theory, the loanable funds theory, Say's law and the dichotomy between the real and the monetary sectors of the economy were discarded in the rebottling process. The resulting model also differed from the traditional classical ideas by being an integrated macroeconomic framework.

The classical paradigm was, in general, rejected by the majority of the economics profession from the 1940s to the 1970s, though it continued to exist as an outcast. However, refinements and additions to it continued to be made during these decades. The dominant paradigm in these decades was the Keynesian one. The classical paradigm, though with new models, roared back in the 1970s and has since then taken various forms. These are the 1970s monetarism, the modern classical model and the new classical model.

III. 1970s monetarism

The *1970s monetarist approach*, also known as the *St Louis monetarism*, was the name given to a mainly empirical analysis whose empirical and theoretical expositions were initiated

30 The analysis of aggregate demand requires the concept of the multiplier, which was proposed only in the 1930s.

by economists at the Federal Reserve Bank of St Louis during the 1970s. The short-run version of their model did not assume full employment and did not imply continuous full employment in the economy. It was relatively close to the then Keynesian models in terms of the impact of monetary policy on output and employment, but it denied on empirical grounds the Keynesian claim of the efficacy of fiscal policy. In its long-run version, it belonged in the classical paradigm.

Therefore, the 1970s monetarism was a hybrid between the classical and the Keynesian paradigms, and made the switch away from Keynesianism palatable for many economists. However, it did not propose any fundamentally new theories, had a short life and was replaced in the early 1980s by ideas truer to the classical paradigm, which eventually took the form of the modern classical paradigm.

IV. Modern classical model

The modern classical model is a statement of the classical paradigm under the assumptions, among others, of *continuous* labor market clearance even in the short-run, which had strictly not been part of the neoclassical model. In addition, for the short-run, this approach extends the neoclassical model by the introduction of uncertainty and rational expectations. In many respects, the modern classical approach is closer to the Walrasian general equilibrium model than to the traditional classical and neoclassical approaches. It is currently the dominant component of the classical paradigm. Its foundation was laid during the 1970s and 1980s.

For the long-run, the modern classical model extends the definition of the (analytical) long-run to include, in addition to the absence of any adjustment costs and rigidities, the assumption that there are no errors, even random ones, in expectations, which is tantamount to the assumption of certainty. Given labor market clearance, this long-run state is the full-employment one.

For the short-run, the modern classical model allows uncertainty, but with expectations formed according to the rational expectations hypothesis. A discussion of the modern classical model appears in Chapter 14. Among its major implications is that deviations from full employment will occur if the expected price level is different from the actual one, so that there are errors in expectations. However, these errors will be random, and by their very nature are transient and self-correcting, so that the short-run deviations from full employment will be *transient and self-correcting*. In this context, systematic monetary and fiscal policies do not change output and unemployment in the short, as well as the long, run. Further, there is no need for such policies since the economy has the ability to go to full employment on its own and within a short period.

Note that, because of the assumption of continuous labor market clearance both in the short-run and in the long-run, involuntary unemployment[31] cannot exist in the modern classical model, even when there are short-run deviations of employment from the full-employment (long-run equilibrium) level.

The modern classical model has serious limitations. In particular, it does not offer a satisfactory explanation for the short-run stylized facts (listed later in Section 1.12) on the impact of shifts in monetary policy on output (see Chapter 14).

31 Involuntary unemployment requires that labor demand exceeds labor supply at the given wage.

Briefly, for the long-run, the modern classical model is a compact form of the Walrasian general equilibrium model, so that its implications are consistent with those of the latter. It provides the benchmark conclusions, consistent with the stylized facts, on the long-run relationship between money and output. For the short-run, the modern classical model produces transient and self-correcting deviations from full employment, so that there is no sensible role for systematic monetary and fiscal policies in both the short-run and the long-run. For the short-run, the implications of the model for output and unemployment are not valid.

V. New classical model

The new classical model imposes the assumption of Ricardian equivalence on the modern classical model. This assumption is an aspect of intertemporal rationality and the Jeffersonian (democratic) notion that the government is nothing more than a representative of its electorate and is regarded as such by the public in making the decisions on its own consumption. Such a government is taken to provide just the goods that the population wants and its bonds, held by the public, are regarded by it (the public) as a debt owed by the public to itself. The implications of these assumptions are that the public debt is not part of the net worth of the public and that the public increases its private saving by the amount of a bond-financed government deficit. The latter implies that such deficits do not affect aggregate demand in the economy, and therefore do not change nominal or real GDP (see Chapter 14 for this analysis).

Of all the macroeconomic models in the classical paradigm, the new classical model is the most restrictive one because of its assumption of Ricardian equivalence.

The major alternative to the classical paradigm is the Keynesian one, which has its own set of models.

1.11 The Keynesian paradigm and the Keynesian set of macroeconomic models

Using the analogy between the economy and the human body

The fundamental difference between the classical and Keynesian paradigms is that while the former focuses on the healthy state of the economy,[32] the latter focuses on the pathology – especially the system-wide pathology – of the economy,[33] which may not fully or soon recover[34] from a shock to it (Solow, 1980, 1991). The Keynesian paradigm recognizes that the economy may sometimes have equilibrium in all markets, but does not assert that this occurs always or most of the time. Further, even if there is equilibrium, it may not be the competitive equilibrium of the Walrasian general equilibrium model because the economy may have a different structure or because of group behavior. As a consequence, the Keynesian paradigm implies that when the economy is outside the Walrasian general equilibrium, the government and the central bank may be able to improve on its actual performance through their policies.

32 That is, with clearance of all markets.
33 That is, when the economy is thrown out of equilibrium.
34 That is, return to equilibrium in all markets.

We have at various places drawn an analogy between the equilibrium state of the economy and the healthy state of the human body, and that between the deviations from equilibrium and the pathology of the human body. The human body sometimes functions in perfect health and sometimes suffers minor illnesses of a brief expected duration and without any need for the help of a professional (doctor). But it could sometimes suffer from serious illnesses from which the recovery may occur but be slow and be speeded up by the help of a doctor, or suffer ones from which there is no recovery without the intervention of a specialist. There may also be illnesses from which there is no cure and no recovery, but we do not include this limiting state within our analogy. Among the serious illnesses, we note there can be many possibilities: infection with bacterium A rather than B, infection by a bacterium versus a virus, an infection versus a collapse of a lung, a collapse of a lung rather than a heart attack, etc. The list of the possible sources of the deviations from the healthy state can be endless.

Comparing the approach of the two paradigms to the pathology of the economy and applying our analogy, when the classical paradigm does envisage deviations away from the healthy state of the economy, they are supposed to be *minor, transitory* and *self-correcting*. Under it, while the economic body may become ill (that is, deviate from the full-employment state), the illnesses are never serious or long lasting, so that a trip to a doctor either never becomes necessary or will not really be worth the hassle and the cost. By comparison, the Keynesian paradigm envisages the possibility of more serious departures from the general equilibrium (healthy) state of the economy. Its deviations from equilibrium can be due to different pathogens or breakdowns of the different components of the economy. Further, it allows for the possibilities that the recovery may be slow and could be speeded up with expert help (from the government and the central bank), or that it may never occur without such help.

Using the analogy with the human body, we offer the following two fundamental – and highly plausible – axioms on the performance of the macroeconomy.

> α. *The economy, like the human body, may sometimes function well and sometimes not.*

Hence, it is essential to study both states, with the former serving as the benchmark for the treatment of the latter.

> β. *When the economy, just like the human body, is not functioning properly, the causes, symptoms and effective treatments of the malfunction can be quite varied.*

The justification for the β axiom is that one cannot plausibly attribute all possible illnesses to a single underlying cause or attribute all potential causes to an overarching single source. An implication of the β axiom is that since the Keynesian paradigm focuses on the pathology of the economy, it cannot properly be encapsulated within one model with one root pathogen. Hence, more than the classical paradigm and its models, which are almost linear or hierarchical in their relationship, the Keynesian paradigm, if it is to do its job properly, has to be a disparate and, at best, a rather loose collection of models.

To reiterate, by the nature of their attempts to deal with the pathology of the economy, the Keynesian models have to be, and are, quite varied. If they are to do their job properly of dealing with the different types of deviations, such models need not – in fact, must not – all focus on the same types of deviation from the overall equilibrium state or make the same recommendations for policies to address these deviations. Unfortunately, this aspect of the Keynesian paradigm is often not recognized. Frequently, the presentations and discussions

of the Keynesian models miss this requirement for variety within the Keynesian paradigm and seek to force the various Keynesian models into a single format or view it as one unified model. The danger in doing so is that a single prescription could be given as a cure-all for very disparate causes and be inappropriate for many.[35] Chapter 15 provides a small number out of the variety of Keynesian models in the literature.

Frequent themes in the Keynesian models

A common concern of the Keynesian models is with the potential for involuntary unemployment, which produces deviations of actual employment from its full-employment level. Consequently, these models tend to pay special attention to the structure of the labor market, its demand and supply functions and whether or not equilibrium holds between them. Within this focus, many Keynesian models assume nominal wage rigidity, often justified by theories of nominal wage contracts between the workers and the firms. However, there are also Keynesian models that consider the deviations from general equilibrium that could occur even when the nominal wage is fully flexible.

The assumption of the rigidity or stickiness of prices in the economy is often regarded as another common theme of Keynesian models. While this assumption can impose deviations from a general equilibrium, it need not be the only cause of or reason for potential deviations. Therefore, models within the Keynesian paradigm need not, and should not, all be based on price rigidity. There is, consequently, also a place for Keynesian models that consider the deviations from general equilibrium that could occur even when the prices are fully flexible.

Chapter 15 provides a look at some of the Keynesian models. While some of the models presented there assume equilibrium in the macroeconomic models, others do not do so. While some assume a special form of the labor supply function, others assume a different form. While some assume – or imply on the basis of nominal wage contracts – nominal wage rigidity of some form, others do not do so. Similarly, while some models assume or imply price level stickiness or rigidity, others do not do so. This variety in modeling within the Keynesian paradigm becomes even more evident when the Keynesian and the neoKeynesian models are compared.

To reiterate, the variety of modeling, though perplexing and sometimes seemingly contradictory, in the Keynesian paradigm is essential to the proper study of the pathology of the economy. It would be a mistake to force the Keynesian models into a single straightjacket, even though this would provide an attractive means of comparing the classical and Keynesian paradigms as a whole.

1.12 Which macro paradigm or model must one believe in?

While most textbooks and economists would consider this to be a legitimate question, our remarks above suggest that it is an improper, and quite likely a dangerous one, for the

35 An example of this is the economists' inappropriate policy prescriptions, based mainly on traditional classical ideas, during the early stages of the Great Depression in the 1930s. These worsened the depth of the fall in GDP and lengthened the depression – and contributed to the demise of faith in the traditional classical ideas. Another example of inappropriate policies, based on the aggregate demand management approach in the Keynesian paradigm, occurred in response to the supply shocks of 1973 and 1974. This led to stagflation and contributed to the demise of faith in the Keynesian paradigm.

formulation of economic policies. The proper study of the economy requires the study of both its healthy state and its diseases. Since we cannot be sanguine that the economy will always operate in general equilibrium, the models of the Keynesian paradigm must not be neglected. Since we cannot be sure that the economy will never be in general equilibrium, the models of the classical paradigm must also not be neglected. Both paradigms have their relevance and usefulness. Neglecting either of them can lead to erroneous policies that impose high costs on the economy and its citizens.

For the practical formulation of monetary policy, the relevant and "interesting" question is not the a priori choice between the classical and the Keynesian models, but rather the perpetually topical one: *what is the current state of the economy like and which model is most applicable to it?* There is rarely a sure answer to this question. Consequently, the judgment on this question and the formulation of the proper monetary policy are an art, not a science – and very often rest on faith in one's prior beliefs about the nature of the economy.

While one cannot dispense with one's beliefs and economists rarely give up their conception of the nature of the economy, the fundamental role of economics must be kept in mind. This is that economics is a positivist science, with the objective of explaining the real world. This is done through its theories, which, by their very nature, must be simplifications – more like caricatures – of reality. As such, they may be valid or not, or be better for explaining some aspects of reality rather than others. Intuition and econometrics are both needed and useful in judging their validity and relative value. In brief, one should not hold a dogmatic belief in one theory for all purposes.

A side implication of the positivist objective of economics is the normative one – i.e. the ability to offer policy prescriptions to improve on the performance of the economy, hopefully as a means of increasing the welfare of its citizens. Both the Keynesian and the classical paradigms are essential to these roles.

One way of judging the extent to which the macroeconomic theories are valid or applicable from a monetary perspective is to compare their implications with the stylized facts of the economy.

Some stylized facts on money and output

Stylized facts on the relationship between money and output are general conclusions about this relationship, established on the bases of intuition and empirical studies. Some of these are:

1. Over long periods of time, there is a roughly one-to-one relationship between the money supply and the price level.
2. Over long periods of time, the relationship between inflation and output growth is not significant.
3. Over long periods of time, the correlation between money growth rates and nominal interest rates is very high.
4. Changes in money supply and interest rates have a strong impact on aggregate demand.
5. Over short periods (a few years), increases in aggregate demand, because of increases in money supply or reductions in interest rates, increase output. This effect builds to a peak and then gradually decreases, so that there is a "hump-shaped pattern" of the effect of monetary policy on output, with the maximum increase in output occurring with a lag longer than one year, sometimes two or more years.

6 The impact of an expansionary monetary policy on prices occurs with a longer lag than on output, so that the impact of monetary shocks on output does not mainly occur through price movements.
7 Contractionary monetary policies initially reduce output significantly, often for longer than a year and sometimes for several years. The cost in terms of output tends to be larger if inflation is brought down gradually rather than rapidly. It is lower if the policy has greater credibility.

Using analytical terminology, money is not neutral in the short-run but is neutral in the long-run. These conclusions hold for monetary policy, whether it changes the money supply or interest rates. Chapter 14 provides a more detailed list of the stylized facts on the impact of monetary policy on output.

1.13 Walras's law

For the closed economy, the standard models of the two paradigms assume four goods: commodities, money, bonds (i.e. all non-monetary financial assets) and labor. Therefore, there should be four equilibrium statements, one for each of the four goods, and the corresponding four curves in the diagrammatic expositions. However, *Walras's law* (see Chapter 18) ensures that equilibrium in any three out of the four markets implies equilibrium in the fourth one, so that one of the markets need not be explicitly studied. This allows the diagrammatic exposition to work with only three equations/curves. Current macroeconomic analysis usually does so for those of the commodity market (the IS equation/curve), the money market (the LM equation/curve if money supply is the instrument of monetary policy but the IR equation/curve if the interest rate is the instrument of monetary policy) and the aggregate supply function (AS equation/curve) or, in its place, a price–output adjustment equation, as in Chapters 14 and 15. In this procedure, the bond market is the one excluded from explicit analysis, so that the bond market curve is not usually drawn. It does, however, remain implicitly in the exposition and can be deduced from the other curves.[36]

1.14 Monetary policy

The standard assumption of monetary analysis was that the central bank exercises control over the economy by exogenously controlling the money supply. In this case, the appropriate analysis of aggregate demand is called IS–LM analysis, since the analysis of the money market generates the IS equation/curve. However, for certain types of economies, controlling the economy's interest rate may be a surer way of controlling aggregate demand than its money supply. The central banks of several developed economies, including those of the United States, Canada and Britain, now seem to rely more on the interest rate rather than on the money supply as the primary monetary policy instrument.[37] For their economies, the LM curve is not appropriate. Instead, the analysis generates an IRT (interest rate target) curve, which, in addition to the IS curve, determines the aggregate demand in the model. The IS–LM and IS–IRT analyses are set out in Chapter 13.

36 This is done in Chapter 19.
37 This is also so for the European Central Bank, which claims to treat the interest rate as its primary monetary policy instrument but also monitors monetary aggregates.

If the central bank sets the interest rate as its exogenous monetary policy instrument, it must be willing to supply the amount of money demanded at that interest rate. It can do this by appropriate changes in the monetary base, either of its volition or by allowing commercial banks to borrow from it. In this case, the money supply becomes endogenous to the economy.

1.15 Neutrality of money and of bonds

Neutrality of money (and credit/bonds) is the proposition that changes in the money supply and monetary policy do not alter output and employment, as well as the real values of many other real variables. For the short run, most models do not imply neutrality. However, as Chapters 13 to 15 show later, the reasons for such non-neutrality differ between the two paradigms and often also among the models of each paradigm. Note that in the long-run analyses of most models, whether in the classical or the Keynesian paradigm, money and credit are neutral, which is consistent with the stylized facts on the economy set out in Section 1.12 and also in Chapter 14.

Money and credit (non-monetary financial variables) are usually not neutral in the *short term* in real-world economies. Sudden shifts in the availability of money and credit are among the most important reasons for fluctuations in output and unemployment. Notable examples of such non-neutrality are provided by currency, credit and exchange crises, which originate in the financial sector and spread to the real sectors of the economy.

An illustration: the subprime crisis of 2007 in the USA

The "subprime crisis" originating in the United States in 2007, and its impact on the real sectors of the US and world economies, provide a compelling illustration of the non-neutrality of both money and credit in the economy. Subprime loans in this context were loans made as mortgages to borrowers who were poor credit risks in terms of their incomes and the collateral that they could provide. However, when house prices were rising sharply, such mortgages seemed to be a good bet for both borrowers and lenders. House prices rose sharply from 2002 to 2006, at some point becoming a "bubble."[38] These mortgages were bundled into "asset-backed corporate securities," which were sold in financial markets and held by a wide variety of financial firms, especially investment bankers, both in the USA and in other countries. These securities were used, in turn, to back up short-term commercial securities sold by financial firms to corporations as liquid, safe investments. As the bubble in US house prices began to collapse in 2006 and house prices fell, the concern over defaults by mortgagees sharply reduced the demand for mortgage-backed corporate securities, as well as the funds made available for loans in this market.[39] This process also increased the general awareness of risk and the risk premium – labeled as the re-pricing of risk – for other types of bonds, so that the ability of households and firms generally to obtain funds for their expenditures became curtailed and the cost of external

38 Prices are said to have a bubble if they exceed the price implied by the fundamentals of demand and supply in the market.

39 The securities backed by risky mortgages are very small compared with the financial assets of banks and other economic agents, but the uncertainty about how much of such securities is held in a particular firm's portfolio creates a hidden risk and increases the risk to lenders of providing further credit to it.

funds increased.[40] These made it difficult for households to buy houses,[41] as well as making it difficult for some corporations to finance their short-term operations,[42] which threatened to reduce production and force the US economy into a recession. The US Federal Reserve System and the European Central Bank, as well as the central banks in many other countries, reacted to the crises in the credit markets by measures to substantially increase the money supply, as well as by reductions in interest rates. In August 2007, while there was considerable uncertainty in the impact of the subprime crisis in financial markets on the real sectors of the economy, there was a general consensus among economists, market analysts, governments and central bankers that, barring appropriate and aggressive monetary policies, the financial crisis would result in a recession in the United States and that this would spread to the world economy.

The impact of the subprime crisis on economic activity, the monetary responses to it and the assessments of the economics profession, as well as those of central bankers and others, clearly show that:

- The consumption and production sectors of the economy depend vitally on the credit sector, so that the supply of credit in the economy is not neutral.
- The supply of credit is not independent of the money supply and interest rates, which are the instruments of monetary policy, so that monetary policy is also not neutral.

To conclude, realistic short-run models of the economy need to embody assumptions about the credit and money markets, and the links between them and consumption and production sectors, that are necessary to imply such non-neutrality. However, few do so. Chapter 16 does so by embodying a link between the supply of short-term loans for working capital and production, as well as a link between such loans and the money supply.

1.16 Definitions of monetary and fiscal policies

The major policy concern of monetary economics is with the impact of monetary policies on the economy. Monetary policy is defined as policy-induced changes in the money supply or/and in interest rates. The control of monetary policy will be taken to be by the central bank or the monetary authority, using these terms as synonymous. The Walrasian general equilibrium and the modern classical models (Chapter 14) imply that, even in the short-run, there is no positive benefit in terms of higher output or lower unemployment from their systematic or anticipated operation (Friedman, 1977; Lucas, 1996), though there are short-run transient effects of random policies. The Keynesian models usually imply that there are such benefits in the short run.

40 This occurred not only in the USA but also in many European, and other, countries because banks and corporations in those countries either held US subprime mortgage-backed securities or because of contagion, which made them reassess the riskiness of their portfolios and also raise their premium for risk.

41 A decline in house construction due to a decline in the demand for housing, when the availability of mortgages fell, was an immediate result.

42 The problem was not that the corporations, which issue commercial paper to fund their day-to-day operations, became less credit-worthy but that the fear of shaky mortgages in their portfolios made investors, including banks, wary of all commercial paper. Fears that banks' own holdings of commercial paper, backed by the mortgage-backed securities, damaged their solvency and profitability even made banks more reluctant to lend to each other.

Fiscal policy is the use of government expenditures, taxes and deficits (or surpluses) as a policy to change the economy. While government deficits can be financed through increases in the money supply (and surpluses be accompanied by decreases in it), macroeconomics defines fiscal policy as one in which the money supply is held constant, so that the deficits must be financed by government borrowing through increases in its bonds sold to the public. Similarly, fiscal surpluses are assumed to require purchases of bonds from the central bank and their retirement, without changing the money supply in circulation in the economy. The reason for this definition of fiscal policy is to separate the effects of changes in the fiscal variables from those in the money supply. To reiterate, fiscal policy is, by definition, *bond-financed fiscal policy*.

In the real world, fiscal and monetary policies are intertwined, more so in some countries than others. However, for analytical purposes, they have to be treated as conceptually independent ones. Hence, a money-financed expansionary fiscal policy – that is, deficits financed by increases in the money supply – will be treated as having two components: an expansionary (bond-financed) fiscal policy and an expansionary monetary policy.

Conclusions

Money performs the two main functions of medium of payments and store of value, with the former being absolutely critical to the transactions role of money in the economy. These functions are performed by a variety of assets, with their liquidity characteristics and substitutability among them changing over time. Innovations in the types of assets and the changing characteristics of existing financial assets mean that the financial assets which meet the role of money keep changing over time.

While currency was considered to be the only form of money at one time, currency and demand deposits were taken to be the only components of money early in the twentieth century, so that the appropriate measure of money was considered to be M1. By 1960, the measure of money had expanded to include time and savings deposits in commercial banks, and therefore had become M2. In subsequent decades, as the liabilities of near-banks became more and more similar to the demand and time deposits of banks, the measures of money were broadened to include the deposits in near-bank financial intermediaries.

The recent incursion of electronics into banking in the form of automatic tellers, banking from home through one's computer or telephone, and the use of smart cards for payments, etc., represents a very fast pace of technical change in the banking industry. It is a safe bet that the empirically appropriate measure of money is changing and will keep changing in the future. During this period of change, the demand functions for money have tended to become unstable, more so for some definitions than others, so that disputes about the proper measure of money have expanded beyond the simple sum aggregates of M1and M2 to encompass more complex forms.

This chapter has also provided an introduction to the two major paradigms in macroeconomics, classical and Keynesian. Each consists of several models. The classical paradigm usually focuses on the general equilibrium of the economy and its models are closely related to each other. The Keynesian one focuses on the deviations from the general equilibrium of the economy. Since there can be many different causes of such deviations in real-world economies, the Keynesian models are a much more diverse group than the classical ones. Knowledge of both paradigms is essential for the proper understanding of the economy and for the appropriate formulation of monetary policies.

The IS–LM mode of macroeconomic analysis is a mode of exposition of the determination of aggregate demand in models of the classical paradigm, as well as in models of the Keynesian paradigm. However, the IS–LM technique of analysis is inappropriate for economies in which the central bank sets the interest rate, rather than the money supply, in its attempts to control aggregate demand in the economy. This is now the practice of many central banks. In this case, aggregate demand is determined by the IS equation and the interest rate set by the central bank.

Summary of critical conclusions

- The appropriate definition of money keeps changing. There are currently several definitions of money in common usage. These include M1, M2 and broader monetary aggregates.
- All definitions of money include currency in the hands of the public and demand/checking deposits in commercial banks.
- Banks are one type of financial intermediaries but differ from others in that their liabilities in the form of checking and savings deposits are the most liquid of all assets in the economy.
- Financial assets are created, so that an unregulated financial system tends to create a multiplicity of differentiated assets.
- The two main paradigms for macroeconomics are the classical and the Keynesian ones.
- The classical paradigm focuses on the general equilibrium of the competitive economy.
- The Keynesian paradigm focuses on the deviations from the general equilibrium of the competitive economy. There can be a variety of reasons for such deviations, requiring different models for their explanations.
- IS–LM analysis assumes that the central bank uses the money supply rather than the interest rate as the monetary policy instrument and sets its level exogenously. However, the LM equation/curve, and therefore the IS–LM analysis, is inappropriate for the macroeconomic analysis of economies in which the central bank sets the interest rate exogenously. The more appropriate analysis for such economies is the IS–IRT one.
- In the short-run, money and credit are not neutral in real-world economies. They are neutral in the analytical long-run.

Review and discussion questions

1. What are the different ways of defining money in your economy? Compare these with the monetary aggregates commonly used in another selected country. Explain their differences and the reasons for such differentiation.
2. Can banks create money? How and under what conditions? How do banks differ from other financial intermediaries and why do central banks regulate more closely the operations of banks?
3. Why do we observe a wide variety of checking and savings accounts, rather than just one of each type?
4. What are the reasons for the existence of financial intermediaries? Why do the ultimate lenders usually not lend directly to the ultimate borrowers?
5. What are the underlying themes (or theme, if only one) of the classical paradigm? How are they represented in the different models within this paradigm?
6. Explain the various models within the classical approach and compare them. Which would you accept for your economy?

7. Explain Say's law and provide its justification. Discuss its validity for a monetary economy that has commodities, money, and bonds.
8. "The modern classical approach does not assume full employment. In fact, it allows for the deviations of employment from its full-employment level." Discuss these statements. If you agree with them, what is the nature of such deviations? Compare their nature with the nature of deviations from full employment that can occur in the traditional classical and neoclassical approaches and in the 1970s monetarist doctrines.
9. What are the underlying themes of the Keynesian paradigm? Do they justify the study of just one model, one variety of models, or several different varieties of models? Why?
10. In order to explain the performance of the economy through the business cycle and the formulation of the appropriate monetary policy, would you rely on either the classical paradigm or the Keynesian one, or sometimes on one and sometimes on the other? Explain your answer with reference to the different phases of the business cycle.
11. Even if it is assumed that the central bank holds the money supply exogenous, why is it inappropriate to use the IS–LM equations/curves only for the determination of real output for both the closed and open economies? Frame your answer in terms of the implications of Walras's law.
12. For a designated country of your choice, what is the appropriate assumption for macroeconomic analysis on the exogeneity or endogeneity of the money supply? What justifies this assumption?
13. What aspects of the economy should the central bank examine in making its decision on whether to use the money supply or the interest rate as its primary/exogenous monetary policy instrument?
14. Why is the IS–LM analysis inappropriate for an economy in which the central bank sets the interest rate exogenously? How would the money supply be determined in this context?
15. "The 1970s monetarism was a hybrid between the classical and the Keynesian paradigms." Discuss.
16. "Under the modern classical approach, there is no sensible role for demand management policies in both the short-run and the long-run." Why not? Discuss.

References

Friedman, M. "Nobel prize lecture: inflation and unemployment." *Journal of Political Economy*, 85, 1977, pp. 451–73.

Goodhart, C.A.E. *Monetary Theory and Practice. The UK Experience.* London: Macmillan, 1984.

Lucas, R.E., Jr. "Nobel lecture: monetary neutrality." *Journal of Political Economy*, 104, 1996, pp. 661–82.

Radford, R.A. "The economic organisation of a P.O.W. camp." *Economica*, 12, 1945, pp. 189–201.

Solow, R.M. "On theories of unemployment." *American Economic Review*, 70, 1980, pp. 1–11.

Solow, R.M. "Cowles and the tradition of economics." In *Cowles Fiftieth Anniversary, Four Essays and an Index of Publications*. Cowles Foundation, 1991, pp. 81–104.

2 The heritage of monetary economics

The heritage of current monetary theory lies in two different sets of ideas: the classical and the Keynesian. This heritage includes both the microeconomic and macroeconomic aspects of monetary economics.

The monetary aspects of the traditional classical approach were encapsulated in the quantity theory for the determination of the price level and the loanable funds theory for the determination of the interest rate. The statement of the quantity theory was an evolutionary one, with several – at least three – quite distinct approaches to the role of money in the economy. These quite diverse approaches shared the common conclusion that, in equilibrium, changes in the money supply caused proportionate changes in the price level but did not change output and unemployment in the economy. One of these approaches, provided by Knut Wicksell, proved to be a precursor of several aspects of the Keynesian macroeconomic approach.

The Keynesian approach discarded the quantity theory and integrated the analysis of the monetary sector and the price level into the complete macroeconomic model for the economy. For the monetary sector, it elaborated on the motives for holding money, leading to the modern approach to the analysis of the demand for money.

Key concepts introduced in this chapter

- An identity versus a theory
- Quantity equation
- Quantity theory
- Wicksell's pure credit economy
- Transactions demand for money
- Speculative demand for money
- Precautionary demand for money
- Transmission mechanism
- Direct transmission mechanism
- Indirect transmission mechanism
- Lending channel
- Permanent income

The discussion of the role of money in the determination of prices and nominal national income in the economy has chronologically an extremely long heritage, extending back to Aristotle in ancient Greece, with explicit formulation of theories on it emerging in the

mid-17th century. Current monetary theory has evolved from two different streams: the quantity theory stream, which was a part of the classical set of ideas, and the Keynesian one. This heritage includes both the microeconomic and macroeconomic aspects of monetary economics.

The quantity theory is the name given to the ideas on the relationship between the money supply and the price level from the middle of the eighteenth century to the publication of Keynes's *The General Theory* in 1936. It was a fundamental part of the traditional classical approach in economics. The specification of the quantity theory was an evolutionary tradition with several – at least three – distinct approaches to the role of money in the economy. These quite diverse approaches shared the common conclusion that, in long-run equilibrium, the changes in the money supply caused proportionate changes in the price level but did not change output or unemployment in the economy. The three approaches to the quantity theory are those based on the quantity equation (see Fisher's [1911] version of this approach below), on the demand for money in the Cambridge (UK) tradition (see Pigou's [1917] version of this approach below) and on a broader macroeconomic analysis (see Wicksell's [1907] approach below). Of these, the demand-for-money approach led to Keynes's elaboration of money demand, and Wicksell's approach led to both Keynes's and the current new Keynesian macroeconomic determination of the price level in a general macroeconomic framework.

The Keynesian approach discarded certain aspects of the quantity theory ideas and developed others in a new and distinctive format. On the demand for money, it elaborated on the earlier Cambridge approach and also rearranged its presentation in terms of the motives for holding money. This treatment in terms of motives eventually led to the modern treatment of the demand for money in terms of four motives: transactions, speculative, precautionary and buffer stock. The Keynesian emphasis on money as an asset, held as an alternative to bonds, also led to Friedman's analysis of the demand for money as an asset, thereby bringing this approach to money demand into the folds of the classical paradigm. At the macroeconomic level, Keynesian analysis made commodity market analysis, based on consumption, investment and the multiplier, a core part of macroeconomics. In doing so, it followed Wicksell. The Keynesian approach also integrated the analysis of the monetary sector into the complete macroeconomic model for the economy.

This chapter's very brief review of this heritage covers the contributions of David Hume, Irving Fisher, A.C. Pigou and Knut Wicksell for the classical period in economics and of John Maynard Keynes and Milton Friedman for the post-1936 period. In the evolution of ideas, the theoretical and empirical analysis of the demand for money only emerged during the twentieth century as a major element of monetary economics. This chapter reviews the three approaches to the quantity theory, followed by the contributions of Keynes and Friedman on the demand for money. It ends with the review of the transmission channels through which changes in the money supply affect aggregate demand and output.

2.1 Quantity equation

Any exchange of goods in the market between a buyer and a seller involves an expenditure that can be specified in two different ways.

A. Expenditures by a buyer must *always* equal the amount of money handed over to the sellers, and expenditures by the members of a group which includes both buyers and sellers must *always* equal the amount of money used by the group, multiplied by the

number of times it has been used over and over again.[1] Designating the average number of times money turns over in financing transactions as its velocity of circulation V, expenditures as $\$Y$ and the money stock in use as $\$M$, we have $\$Y \equiv \$MV$, where $\equiv$ indicates an *identity* rather than merely an equilibrium condition.

B. Expenditures on the goods bought can also be measured as the quantity of physical goods traded times the average price of these commodities.[2] Expenditures Y then always equal the quantity y of the goods bought times their price level P, so that $\$Y \equiv \$Py$.

Obviously, these two different ways of measuring expenditures must yield the identical amount. These two measures are:

$$Y \equiv MV$$

$$Y \equiv Py$$

Hence,

$$MV \equiv Py \tag{1}$$

where:

y = real output (of commodities)
P = price level (i.e. the average price level of commodities)
Y = nominal value of output ($\equiv$ nominal income)
M = money supply
V = velocity of circulation of money (M) against output (y) over the designated period.

Equation (1) is an identity since it is derived solely from identities. It is valid under any set of circumstances whatever since it can be reduced to the statement: in a given period, by a given group of people, expenditures equal expenditures, with only a difference in the computational method between them. (l) is *true* for any person or group of persons.[3] If it is applied, as it usually is, to the aggregate level for the whole economy, the two sides of the identity and its four variables refer to all expenditures in the economy. But if it is applied to the world economy as a whole, its total expenditures and the four variables will be for the world economy.

(1) is called the *quantity equation*, the word "equation" in this expression serving to distinguish it from the *quantity theory,* which is vitally different in spirit and purpose from the quantity equation. As we shall see later, the quantity theory is not an identity, while the

1 Thus a person buying $100 of goods pays $100 to seller 1. Suppose the latter in turn buys $100 worth of goods from another seller (seller 2) of goods. The total expenditure was thus $200, the amount of money used was only $100 and it was paid over twice in financing the expenditures. Suppose now that the initial seller had bought only $50 worth of goods from seller 2. Total expenditures would now be $150; the amount of money in use remains at $100 but it has been paid over only 1.5 times on average.

2 Since the goods traded are generally of different kinds, there are obviously problems in thinking of an aggregate measure of goods in physical terms and of the price level to be associated with a unit of such a conglomerate or composite good. Both the 'quantity' or 'output' y of this good and its average price P must then be thought of as indices.

3 Identities are said to be true or false. By comparison, propositions or relationships about the real world are said to be valid or invalid.

quantity equation is not a *theory* for the determination of prices, incomes or even the velocity of circulation in the economy.

Note that a relationship or statement that is *always* valid under *any* circumstances is said to be an *identity* or *tautology*. Identities generally arise by the way the terms in the relationship are defined or measured. Thus, (1) defines (measured) expenditures in two different ways, once as MV and then as Py, so that (1) is an identity. An identity is different from an equilibrium condition that holds only if there is equilibrium but not otherwise – i.e. when there is disequilibrium. Further, a *theory* may or may not apply to any particular economy in the real world or it may be valid for some states – e.g. equilibrium ones – but not for others, while an *identity* is true (or false) by virtue of the definitions of its variables and its logic, so that its truth or falsity cannot be checked by reference to the real world. A theory usually includes some identities but must also include behavioral conditions – which are statements about the behavior of the economy or its agents – and often also equilibrium conditions on its markets.

Note also that the velocity of circulation V depends on the length of the period of analysis. Since Y is a flow while M is a stock, the longer the period of analysis, the larger will be Y whereas M will be a constant. Therefore, V will increase with the length of the period.

Policy implications of the quantity equation for persistently high rates of inflation

Rewrite the quantity equation in terms of growth rates as:

$$M'' + V'' \equiv P'' + y''$$

where $''$ indicates the rate of change (also called the growth rate) of the variable. This identity can be restated as:

$$\pi \equiv M'' + V'' - Y''$$

where π is the rate of inflation and is the same as P''. This identity asserts that the rate of inflation is always equal to the rate of money growth plus the growth rate of velocity less the growth rate of output. *Ceteris paribus*, the higher the money growth rate, the higher will be the inflation rate, whereas the higher the output growth rate is, the lower will be the inflation rate. Note that velocity also changes over time and can contribute to inflation if it increases, or reduce inflation when it falls.[4]

In normal circumstances in the economy, velocity changes during a year but not by more than a few percentage points. Similarly, for most economies, real output growth rate is usually only a few percentage points. For the quantity equation, we need only consider the difference $(V'' - y'')$ between them. In the normal case, both velocity and output increase over time but the difference in their growth rates is likely to be quite small, usually in low single digits. *Adding this information to the quantity equation* implies that high (high single digits or higher numbers) and persistent (i.e. for several years) rates of inflation can only stem from high and persistent money growth rates. This is particularly true of hyperinflations in which the annual inflation rate may be in double (10 percent or more) or triple (100 percent or more) digits or

4 The spread of banks and automatic teller/banking machines (ATMs) has tended to increase velocity in recent decades.

even higher. Empirically, even at low inflation rates, the correlation between money supply growth and inflation rates over long periods is close to unity.

To reiterate, the source of inflation over long periods is usually money supply growth and the source of persistently high inflation over even short periods is high and persistent money growth rates. Therefore, if the monetary authorities wish to drastically reduce inflation rates to low levels, they must pursue a policy that achieves an appropriate reduction in money supply growth.

2.1.1 Some variants of the quantity equation

There are several major variants of the quantity equation. One set of variants focuses attention on the goods traded or the transactions in which they are traded, so that they modify the right-hand side of (1). The second set of variants imposes disaggregation on the media of payments (e.g. into currency and demand deposits) or changes the monetary aggregate, thereby modifying the left side of (1). We present some forms of each of these variants. The first set of these variants is given by (i) and (ii) below. The second set is given by (iii).

(i) Commodities approach to the quantity equation

One way of measuring expenditures is as the multiple of the amount y of commodities sold in the economy in the current period times their average price level P. Therefore, the quantity equation can be written as:

$$M \cdot V_{My} \equiv P_y \cdot y \tag{2}$$

where:

V_{My} = income-velocity of circulation of money balances M in the financing of the commodities in y over the designated period
P_y = average price (price level) of currently produced commodities in the economy
y = real aggregate output/income in the economy.

(2) is often also stated as:

$$M \cdot V_{My} \equiv Y \tag{3}$$

(3) yields velocity V_{My} as equaling the ratio Y/M.

(ii) Transactions approach to the quantity equation

If the focus of the analysis is intended to be the number of *transactions* in the economy rather than on the quantity of goods, expenditures can be viewed as the number of transactions T of all goods, whether currently produced or not, in the economy times the average price P_T paid *per transaction*. The concept of velocity relevant here would be the rate of turnover per period of money balances in financing all such transactions. The quantity equation then becomes:

$$M \cdot V_{MT} \equiv P_T \cdot T \tag{4}$$

where:

V_{MT} = transactions-velocity of circulation per period of money balances M in financing transactions T
P_T = average price of transactions
T = number of transactions during the period.

To illustrate the differences between y and T and between P_y and P_T, assume that we are dealing with a single transaction involving the purchase of ten shirts at a price of \$10 each. The total cost of the transaction is \$100. Here, the quantity y of goods is 10 and their average price P_y is \$10, while the number of transactions T is one and their average price P_T is \$100.

(iii) Quantity equation in terms of the monetary base

The monetary base[5] consists of the currency in the hands of the public (households and firms), the currency held by the financial intermediaries and the deposits of the latter with the central bank. Since the central bank has better control over the monetary base, which it can manipulate through open market operations, than over M1 or M2, it is sometimes useful to focus on the velocity of circulation $V_{M0,y}$ of the monetary base. This velocity depends not only upon the behavior of the non-banking public but also upon the behavior of firms and financial intermediaries. The quantity equation in terms of the monetary base is:

$$M0 \cdot V_{M0,y} \equiv P_y \cdot y \tag{5}$$

where:

$M0$ = quantity of the monetary base
$V_{M0,y}$ = income-velocity of circulation per period of the monetary base.

The quantity equation is thus a versatile tool. Note that all versions of it are identities. The form in which it is stated should depend upon the analysis that is to be performed. Examples of such interaction between the intended use and the actual variant of the quantity equation employed occur often in monetary economics.

2.2 Quantity theory

The quantity theory had a rich and varied tradition, going as far back as the eighteenth century. It is the proposition that *in long-run equilibrium, a change in the money supply in the economy causes a proportionate change in the price level, though not necessarily in disequilibrium*.

The quantity theory was dominant in its field through the nineteenth century, though more as an approach than a rigorous theory, varying considerably among writers and periods. Two versions of the form that it had achieved by the beginning of the twentieth century are presented below from the works of Irving Fisher and A.C. Pigou. A third version, radically different from those of these writers, is presented later from the writings of Knut Wicksell.

5 The monetary base is also sometimes called the reserve base or high-powered money.

2.2.1 Transactions approach to the quantity theory

Irving Fisher, in his book *The Purchasing Power of Money* (1911), sought to provide a rigorous basis for the quantity theory by approaching it from the quantity equation. He recognized the latter as an identity and added assumptions to it to transform it into a theory for the determination of prices. A considerable part of his argument was concerned with providing a clear and relevant exposition of the quantity equation, and one of his versions of this equation is presented below.

Fisher distinguished between currency and the public's demand deposits in banks. This distinction was relevant to the economy when he wrote, since currency was commonly used in payments whereas payments by check were much less common. Setting aside this distinction for the modern economy, we use M1 as the relevant money variable. Fisher also stated his version of the quantity theory in terms of the number of transactions, rather than in terms of the quantity of commodities purchased.[6] However, as a result of Keynes's emphasis on national income/output rather than total transactions, while data on national income/output came to be gathered and made commonly available, the data on the number of transactions was not gathered and has not become available in the public domain. The following, therefore, adapts Fisher's treatment of the quantity equation and theory and couches it in terms of the amount of the commodities purchased rather than in terms of transactions. This adapted version of the quantity equation has the form:

$$MV \equiv Py \tag{6}$$

To transform the quantity equation into the quantity theory, Fisher put forth two propositions about economic behavior. These are:

> (i) The velocities of circulation of "money" (currency) and deposits depend … on technical conditions and bear no discoverable relation to the quantity of money in circulation. Velocity of circulation is the average rate of "turnover" and depends on countless individual rates of turnover. These … depend on individual habits. … The average rate of turnover … will depend on density of population, commercial customs, rapidity of transport, and other technical conditions, but *not on the quantity of money and deposits nor on the price level*.
> (ii) *(except during transition periods)* the volume of trade, like the velocity of circulation of money, is *independent of the quantity of money*. An inflation of the currency cannot increase the product of farms and factories, nor the speed of freight trains or ships. The stream of business depends on natural resources and technical conditions, not on the quantity of money. The whole machinery of production, transportation and sale is a matter of physical capacities and technique, none of which depend on the quantity of money.
>
> (Fisher, 1911).

6 Fisher's version of the quantity equation is $CV_C + DV_D \equiv P_T T$, where C is currency and V_C is its velocity, while D is demand deposits and V_D is its velocity. Total expenditures equal $(CV_C + DV_D)$. Fisher maintained "that bank reserves are kept in a more or less definite ratio to bank deposits" and "that individuals, firms and corporations maintain more or less definite relations between their money (currency) and deposit balances." Hence, C and D will always change in proportion. The alternative way of measuring expenditures in Fisher's transactions approach would be as $P_T \cdot T$, where P_T is the average price of transactions and T is the aggregate number of all transactions against commodities.

Therefore, Fisher's conclusion was that:

> while the equation of exchange is, if we choose, a mere "*truism*" based on the equivalence, in all purchases, of the money … expended, on the one hand, and what they buy on the other, yet *in view of supplementary knowledge … as to the non-relation of* [*velocity to money and prices*], this equation is the means of demonstrating the fact that normally the *prices vary directly as M*, that is, demonstrating the quantity theory.
>
> (Fisher, 1911, italics and the clause in brackets added).[7]

Fisher was certainly right in specifying that the transformation from his version of the quantity equation to the quantity theory *requires* that, when the monetary authorities increase the amount of money, the velocity of circulation and the quantities of goods remain unchanged. These assertions, as well as (i) and (ii) above, are economic ones, resting on assumptions about human behavior, and may or may not be valid. In symbols and in the above updated mode of statement of the quantity equation, these assertions become: $\partial y/\partial M = 0$ and $\partial V/\partial M = 0$. These imply that, following an increase in the money supply, prices will rise in proportion to the increase in the money supply. That is, the elasticity of the price level with respect to the money supply will be unity.[8]

Fisher pointed out that the above assertions did not necessarily apply during "transition" (which can be interpreted as "disequilibrium") periods, so that his assertions applied to a comparison of the equilibrium states prior to and after a one-time increase in the money supply. Fisher based these assertions on the then dominant theories of output and other real variables (including velocity), for which the traditional classical approach and Walrasian model imply the independence of real variables from the monetary ones, which are M and P, in equilibrium.

On assumption (ii) of Fisher, the dominant theory – which was part of Fisher's own views of the economy – of the early twentieth century on output and employment in the economy was the Walrasian one, which treated each market separately and used microeconomic analysis.[9] This analysis implied that the labor market would clear in equilibrium and there would be full employment. Output would tend to stay at the full-employment level, except in the transient disequilibrium stages. Further, this full-employment output was independent of the money supply and prices. Therefore, Fisher's assertion that changes in the money supply would not affect the equilibrium output of goods was consistent with the real economic theory of the time and was, in effect, based on the latter. This assertion was to be later challenged by Keynes and the Keynesians for demand-deficient economies, reaffirmed by the modern classical economists in the 1980s and 1990s and denied by the new Keynesians in the last two decades.[10] Further, note Fisher's qualification "*except during transition periods*" to the quantity theory proposition. Interpreting this as a reference to the disequilibrium induced by an exogenous change in the money supply, the real-time of this transition (from one long-run

7 The symbols have been changed and italicized in the above quotation as well as in following ones.

8 To derive this from Fisher's arguments, take the derivative, with respect to M, of the quantity equation $MV \equiv Py$, where the symbol Q for real output has been replaced by y. This yields: $V + M \cdot \partial V/\partial M = y \cdot \partial P/\partial M + P \cdot \partial y/\partial M$. Fisher's argument is that, in equilibrium, $\partial V/\partial M = 0$ and $\partial y/\partial M = 0$. Hence, in equilibrium, $(M/P) \cdot (\partial P/\partial M) = 1$, which is the quantity theory proposition.

9 See Chapter 1.

10 See Chapters 14 to 17 for more information on these schools.

equilibrium state to the one following a change in the money supply) becomes a very relevant question for the pursuit, or not, of monetary policy.

Fisher's assumption (i) on the independence of velocities from changes in the money supply is also questionable. The velocity of circulation of money is not directly related to the behavior of firms and households and, if one thinks solely in terms of velocity, Fisher's simplistic argument on this point seems reasonable. However, since velocity is a ratio of expenditures to money holdings, Fisher's assertion becomes more easily subject to doubt if the determinants of velocity are approached from the determinants of expenditures and the demand for money, as Keynesians do, and if the economy is not continuously in general equilibrium at full employment. These determinants include interest rates and output, so that changes in interest rates and in output can change both the demand for money and its velocity.

However, velocity is a real variable since it can be defined as equal to real income divided by the real money stock in the economy. Modern classical economists focus on velocity as a real variable, as Fisher had done, and, along with other real variables, take it to be independent of money supply and the price level in the long-run equilibrium state of the economy. Hence, modern classical economists agree with both of Fisher's assumptions for the general equilibrium – that is, with all markets clearing – state of the economy. Modern classical economists, therefore, *with a model implying continuous full employment*, still maintain Fisher's quantity theory assertion that an increase in the money supply will cause a proportionate increase in the price level, with velocity remaining unchanged.

Keynesians question the empirical usefulness of the assumption of continuous long-run general equilibrium (yielding full employment) since they maintain that continuous full employment does not normally exist in the economy. They also assert the dependence of money demand on the interest rate and the dependence of the interest rate on liquidity preference and the money supply. Hence, to the Keynesians, neither velocity nor output is independent of the money supply. Therefore, Keynesians reject the validity of the quantity theory both in terms of comparison across equilibrium states and in disequilibrium.

Determinants of velocity: constancy versus the stability of the velocity function

Equilibrium in the money market means that money demand equals money supply. Therefore, in this equilibrium, velocity can be redefined as the nominal income divided by money demand. As explained in subsequent sections of this chapter, money demand depends upon many variables, of which the most important are national income and interest rates. As income rises, economies of scale in money holdings mean that money demand does not rise as fast, so that velocity increases. The interest rate is the cost of holding money rather than interest-paying financial assets, so that money demand falls as interest rates rise, which increases velocity. Therefore, velocity rises as income rises and also rises as interest rates rise.

Financial innovations in recent decades have created a variety of substitutes for M1 and M2, which have reduced their demand. Further, telephone and electronic banking have reduced the need to hold large precautionary balances against unexpected needs for expenditures. This trend has been reinforced by the fall in brokerages costs of various types in switching between money and other financial assets, so that individuals can manage their expenditures with smaller money balances while holding larger amounts of interest-paying financial assets. These developments have reduced the demand for M1 and M2, so their velocity has risen

considerably in recent decades. To illustrate, while the velocity of M1 in the USA was about 6.3 in 1991, it rose to about 8.8 in 2000.

Table 2.1 shows that the velocity of circulation, which equals nominal national income divided by the money supply, in Canada, UK and USA is not a constant. In fact, it varies even over periods as short as a day or month.

Fisher did not assume the constancy of velocity. His assumption was the independence of velocity – a real variable – from that of changes in the money supply and the price level in the general equilibrium states of the economy. From an empirical perspective, velocity is not a constant in either the short term or the long term in actual economies. It is continuously changing in the economy. Some estimates of the average annual change in velocity for the USA lie at about 3 percent to 4 percent.

To conclude, Fisher did not assume velocity to be a constant, nor is it constant in the real economy. Economic theory takes it to be an economic variable, determined in the economy by other economic variables. As its determinants change, velocity changes. The determinants of velocity are discussed in greater detail later in this chapter.

Table 2.1 Changes in the velocity of money

	Velocity of M1 *and* M2[11] *for USA (US$ billions)*				
	Y[12]	M1	M2	*V*1 *(Y/*M1*)*	*V*2 *(Y/*M2*)*
1991	5803.075	916.0	3472.7	6.34	1.67
1995	7397.650	1150.7	3680.0	6.43	2.01
2000	9816.975	1112.3	4962.2	8.83	1.98
	Velocity of M1 *and* M2+ *for Canada (C$ billions)*				
	Y	M1	M2+	*V*1*(Y/*M1*)*	*V*2 *(Y/*M2+*)*
1991	679.921	45.622	534.989	14.90	1.27
1995	810.426	62.674	618.447	12.93	1.31
2000	1076.577	114.919[13]	713.503	9.37	1.51
	Velocity of M2 *and* M4 *for UK (British pounds billions)*				
	Y	M2	M4	*V*2 *(Y/*M2*)*	*V*4 *(Y/*M4*)*
1991	558.160	278.3	502.1	2.01	1.11
1995	719.747	437.0	622.6	1.65	1.16
2000	953.227	600.3	885.0	1.59	1.08

11 The data on M1, M2, M2+ and M4 in Table 2.1 are taken from *Statistics on Payment Systems in the Group of Ten Countries*, published by the Bank for International Settlements, various years.

12 The data on nominal GDP (*Y* in Table 2.1) are taken from the *World Economic Outlook Database* of the International Monetary Fund, September, 2006.

13 This figure is taken from *Weekly Financial Statistics* of the Bank of Canada, March 2, 2001.

The Fisher equation on interest rates: distinction between nominal and real interest rates

Another of Fisher's contributions on monetary theory was his distinction between the nominal and real interest rates. This is embodied in what has been designated the Fisher equation.

The rate of interest that is charged on loans in the market is the *market* or *nominal rate of interest*. This has been designated by the symbol R. If the rational lender expects a rate of inflation π^e, he has to consider the real interest rate r that he would receive on his loan. However, financial markets usually determine the nominal interest rate R. In perfect capital markets, the *ex ante* relationship[14] between the expected real interest rate r^e and the nominal interest rate R is specified by

$$(1+r^e) = (1+R)/(1+\pi^e) \tag{7}$$

where π^e is the expected inflation rate. If there exist both real bonds (i.e. promising a real rate of return r per period) and nominal bonds (i.e. promising a nominal rate of return R per period), the relationship between them in perfect markets would be:

$$(1+R) = (1+r)(1+\pi^e) \tag{8}$$

At low values of r^e and π^e, $r^e\pi^e \to 0$, so that (7) is often simplified to:

$$r^e = R - \pi^e \tag{9}$$

This states that the real yield that the investor expects to receive equals the nominal rate minus the expected loss of the purchasing power of money balances through inflation. (8) is correspondingly simplified to $R = r + \pi^e$. (8) and (9) are known as the Fisher equation.

Note that the real value of the rate of return that the holder of a nominal bond would actually (i.e. *ex post*) receive from his loan is the *actual real rate of interest* (r^a), which is correspondingly given by:

$$r^a = R - \pi \tag{9'}$$

In these equations, the definitions of the symbols are:

R = nominal rate of interest
r^a = actual (*ex post*) real rate of interest on nominal bonds
r = real rate of interest
r^e = expected real rate of return
π = actual rate of inflation
π^e = expected rate of inflation.

14 One explanation for the Fisher equation is as follows. An investor investing one dollar in a "nominal bond" (i.e. paying a nominal rate of interest R) would receive $\$(1+R)$ at the end of the period. If he were to buy a "real bond" (i.e. paying a real interest rate r), he would receive $(1+r)$ in real terms (i.e. in commodities) at the end of the period. Given the expectations on inflation held at the beginning of the current period, the expected nominal value at the end of the period of this real amount equals $\$(1+r)(1+\pi^e)$. The investor would be indifferent between the nominal and the real bonds if the nominal return from both bonds were equal, i.e. $(1+R) = (1+r)(1+\pi^e)$. With all investors behaving in this manner, perfect capital markets would ensure this relationship.

If the actual rate of inflation were imperfectly anticipated, the actual yield r^a on nominal bonds would differ from the expected one r^e and may or may not be positive. In fact, negative real interest rates are often observed during years of accelerating inflation, such as in the 1970s, when the real yield on nominal bonds was often, and often persistently, negative.[15]

Fisher's direct transmission mechanism

For the transmission mechanism from exogenous money supply changes to the endogenous changes in aggregate demand and prices, Fisher argued that an increase in the money supply leads its holders to increase their expenditures on commodities. Fisher's version of this disequilibrium chain of causation from changes in the money supply to changes in the nominal value of aggregate expenditures is given in the following quotation. Fisher starts by assuming that an individual's money holdings are doubled, and continues as:

> Prices being unchanged, he now has double the amount of money and deposits, which his convenience had taught him to keep on hand. *He will then try to get rid of the surplus money and deposits by buying goods*. But as somebody else must be found to take the money off his hands, its mere transfer will not diminish the amount in the community. It will simply increase somebody else's surplus.… Everybody will want to exchange this relatively useless extra money for goods, and the desire so to do must surely drive up the price of goods. [This process will continue until prices double and equilibrium is restored at the initial levels of output and velocity.]
>
> (Fisher, 1911, italics added).

Fisher's mechanism, by which changes in the money supply induce changes in aggregate expenditures, has come to be known as the *direct transmission mechanism* of monetary policy, as compared with the *indirect transmission mechanism*, which relies upon the changes in the money supply inducing changes in interest rates, which in turn induce changes in investment, which then cause changes in aggregate expenditures. The latter mechanism was incorporated in the 1930s into the Keynesian and neoclassical macroeconomic models, but the former was revived by Milton Friedman and the 1970s monetarist models. The modern classical models generally ignore the direct transmission mechanism and, as with the Keynesian models, incorporate the indirect transmission mechanism. However, the direct transmission mechanism continues to be relevant to the poor whose expenditures are close to their incomes, and especially in economies in which the increase in the money supply is used to finance fiscal deficits and initially ends up in the hands of people whose usual use of extra funds is to buy commodities.

2.2.2 Cash balances (Cambridge) approach to the quantity theory

Another popular approach to the quantity theory examined the determination of prices from the perspective of the demand and supply of money. Some of the best known exponents of

15 However, under the rational expectations hypothesis, the error in expectations would be only random and uncorrelated with information available at the time the expectations are formed, which implies that the real rates could not be persistently negative.

this approach were at Cambridge University in England and included, among others, Alfred Marshall, A.C. Pigou and the early writings (that is, pre-1936) of John Maynard Keynes. The following exposition of this approach follows that of Pigou in his article, *The Value of Money* (1917).

Pigou, like Fisher, defined currency or *legal tender* as money but was, in general, concerned with what he called "*the titles to legal tender.*" He defined these titles as including currency and demand deposits in banks, which correspond to the modern concept of M1. He argued that a person held currency and demand deposits:

> to enable him to effect the ordinary transactions of life without trouble, and to secure him against unexpected demands due to a sudden need, or to a rise in the price of something that he cannot easily dispense with. For these two *objects*, the *provision of convenience* and the *provision of security*, people in general elect to hold currency and demand deposits.
>
> (Pigou, 1917, italics added).

The actual demand for currency and demand deposits is:

> determined by the *proportion* of his resources that the average man chooses to keep in that form. This proportion depends upon the convenience obtained and the risk avoided through the possession of such titles, by the loss of real income involved through the provision to this use of resources that might have been devoted to the production of future commodities, and by the satisfaction that might be obtained by consuming resources immediately and not investing at all.
>
> (Pigou, 1917).

Pigou thus claimed that the individual is not directly concerned with the demand for money but with its relation to his total resources. These resources can be interpreted as wealth in stock terms or as income/expenditures in flow terms. We will use the latter, so that income will be the proxy for Pigou's "resources." Further, according to Pigou, this ratio of money demand to resources is a function of its services, the internal rate of return on investments and of the marginal satisfaction foregone from less consumption. Representing the internal rate of (real) return on investment as r and assuming it to be an approximate measure, in equilibrium, of the satisfaction foregone by not consuming, the ratio of money balances demanded (M^{d}) to total nominal expenditures (Y) is given by:

$$M^{\mathrm{d}}/Y = k(r) \quad k'(r) < 0 \tag{10}$$

where k is a *functional* symbol. M^{d}/Y decreases with r, or, in Pigou's words, "the variable k will be larger the less attractive is the production use and the more attractive is the rival money use of resources." Hence, $\partial k/\partial r < 0$. Therefore, the demand for money balances, M^{d}, is:

$$M^{\mathrm{d}} = k(r)Y \tag{11}$$

Determination of the price level in the cash balance approach

Assuming a given money supply M, equilibrium in the money market with (11) requires that:

$$M = k(r)Y \tag{12}$$

Writing Py for Y, with P as the price level and y as the real amount of goods,

$$M = k(r)\,Py \tag{13}$$

Assuming that output y is at its full employment level y^{f} in equilibrium, $y = y^{\mathrm{f}}$, so that (11) becomes:

$$M = k(r)\,Py^{\mathrm{f}}$$

where $\partial y^{\mathrm{f}}/\partial P = 0$ and $\partial y^{\mathrm{f}}/\partial M = 0$. Further, Pigou assumed[16] that the equilibrium rate of return (r^*) was determined by the marginal productivity of capital (MPK), which was taken to be independent of the money supply and the price level, so that $\partial r^*/\partial P = 0$ and $\partial r^*/\partial M = 0$. Therefore, in equilibrium,

$$M = k(r^*)\,Py^{\mathrm{f}} \tag{14}$$

so that, in equilibrium,

$$P = M \Big/ \left[k\left(r^*\right) \cdot y^{\mathrm{f}}\right] \tag{15}$$

which implies that:

$$\partial P/\partial M = 1 \Big/ \left[k\left(r^*\right) \cdot y^{\mathrm{f}}\right]$$

and

$$E_{P \cdot M} = (M/P) \cdot (\partial P/\partial M) = 1$$

where $E_{P \cdot M.}$ is the elasticity of P with respect to M. Since this elasticity equals unity, the price level will, in *comparative static equilibria*, vary proportionately with the money supply. Therefore, (14) establishes Pigou's version of the quantity theory proposition.

The cash balance approach starts its statement of the quantity theory as a theory of demand, supply and equilibrium in the money market and then proceeds to place it in a long-run general equilibrium approach to the economy. From a rigorous standpoint, it does not become a theory of the price level until the complete model – which includes the determination of output and interest rates – is specified. On the latter variables, Pigou and his colleagues in the quantity theory tradition had in mind the then generally accepted traditional classical ideas on the determination of output and interest rates. As stated in Chapter 1, these ideas implied the

16 Pigou implicitly did so in *The Value of Money*. This was consistent with his ideas on the determination of the equilibrium rate of return in the economy by the marginal productivity of capital.

independence of the long-run equilibrium values of both these variables from the demand and supply for money and turned the money market equilibrium equation (11) into a statement of the quantity theory. The essential deficiency in Pigou and the cash balance approach lay not so much in their specification of money demand relevant to the time in which they were writing, but in that of the existing (traditional classical) macroeconomic analysis which failed to specify the determination of aggregate demand and its impact on output in short-run equilibrium, as well as in disequilibrium. This deficiency was the major point of attack by Keynes on the quantity theory and the traditional classical approach generally.

Velocity in the cash balance approach

On the velocity of circulation V in Pigou's analysis, we have from (11) that:

$$\begin{aligned} V &= Y/M \\ &= 1/[k(r)] \end{aligned} \tag{16}$$

In (13), since velocity depends upon the rate of interest, it is not a constant in the context of Pigou's money market analysis. However, given the independence of the equilibrium rate of interest and the marginal productivity of capital from the supply of money, the *equilibrium* level of velocity equals $[1/k(r^*)]$, which is independent of the supply of money. This independence of velocity with respect to the money supply does not mean its constancy over time, since velocity could still depend upon other variables, such as banking practices and payment habits, and these often change over time. Further, the independence of velocity from the money supply was asserted only for equilibrium but not for disequilibrium. However, Pigou and other economists in the Cambridge school often fell into the habit of treating k as a constant even though it was a functional symbol with $k'(r) < 0$, so that velocity also became a constant both in and out of equilibrium.

Legacy of the cash balance approach for the analysis of the demand for money

Further developments in monetary theory during the twentieth century built on two aspects of the nineteenth and early twentieth century monetary theory. These were as follows: (i) The cash balance approach started its presentation of the quantity theory by analyzing the demand for money and equilibrium in the money market. This idea was later taken up by Milton Friedman (whose contribution on this topic is presented later in this chapter) to identify and confine the quantity theory to the analyses of the demand for money and the money market; (ii) the cash balance approach had analyzed the demand for money in terms of its characteristics or functions, which were:

1 *The provision of convenience in transactions*.
2 *The provision of security against unexpected demands due to a sudden need or to a rise in the prices.*

The former was related to the demand for the medium of exchange function of money and the latter to its store of value function. These reasons for holding money were restated by Keynes in 1936 into the transactions motive and the precautionary motive. Keynes added to these the speculative motive.

2.3 Wicksell's pure credit economy

Knut Wicksell was a Swedish monetary economist writing within the classical tradition in the last decades of the nineteenth and the first quarter of the twentieth century and considered himself to be an exponent of the quantity theory. His treatment of the quantity theory was very distinctive and quite different from the English and American traditions of the time, as represented in the works of Fisher, Pigou and Keynes during his classical period prior to 1930. Further, elements of Wicksell's analysis led to the formulation of modern macroeconomic analysis. His ideas have assumed even greater importance in the past two decades since several central banks in developed economies have adopted the use of the interest rate as their primary monetary policy instrument, so that the appropriate analysis has to take the interest rate rather than the money supply as being exogenously set. The money supply becomes endogenous in this context. These assumptions are essentially similar to those made by Wicksell. The new Keynesian analysis embodies these assumptions, so that it is sometimes referred to as the neoWicksellian analysis.

Wicksell sought to defend the quantity theory as the appropriate theory for the determination of prices against its alternative, the *full cost pricing* theory. The latter argued that each firm sets the prices of its products on the basis of its cost of production, including a margin for profit, with the aggregate price level being merely the average of the individual prices set by firms. The amount of the money supply in the economy adjusts to accommodate this price level and is therefore determined by the price level, rather than determining it. Wicksell considered this full cost pricing theory as erroneous and argued that such pricing by firms determined the relative prices of commodities, rather than the price level. In his analysis, the latter was determined by the quantity of money in the economy relative to national output since commodities exchange against money and not against each other.

In his reformulation of the quantity theory, Wicksell (1907) sought to shift the focus of attention to the transmission mechanism relating changes in the money supply to changes in the price level. He specified this mechanism for economies using either metallic or fiat money and for a *pure credit* economy. The latter analysis is the more distinctive one and illustrates Wicksell's transmission mechanism more clearly. It is also the one likely to be more relevant to the future evolution of our present day economies and, therefore, is the one presented below.

In modern macroeconomic terminology, Wicksell's analysis of the pure credit economy is essentially short run since his analysis assumes a fixed capital stock, technology and labor force in the production of commodities. This focus on the short run contrasts with Fisher's and Pigou's reliance on the long-run determination of output in order to establish their versions of the quantity theory. Further, Wicksell assumes that the economy is a pure credit one in the sense that the public does not hold currency and all transactions are paid by checks drawn on checking accounts in banks, which do not hold any reserves against their demand deposits. Since the banks do not hold reserves and any loans made by them are re-deposited by the borrowers or their payees in the banks, the banks can lend any amount that they desire without risking insolvency. Further, banks are assumed to be willing to lend the amount that the firms wish to borrow at the specified market rate of interest set by the banks. Wicksell calls the nominal rate of interest at which the banks lend to the public the *money* or *market rate of interest*. The banks accommodate the demand for loans at this interest rate, which is set by them. Under these assumptions, the amount of money supply in the economy is precisely equal to the amount of credit extended by the banks, since these loans are wholly deposited in the banks. Hence, changes in the money

supply occur only when the demand for loans changes in response to an exogenous shift in the interest rate charged by banks. Note that, in Wicksell's pure credit economy, the economy's interest rate is set exogenously by the banks, while the money supply depends on this interest rate and the public's demand for loans. Therefore, it is endogenous to the economy.

A critical element of Wicksell's (1907) theory is the emphasis on saving and investment in the economy. Funds for (new) investment come from saving plus changes in the amount of credit provided by banks. The rate of interest which equates saving and investment was labeled by Wicksell the *normal rate of interest.* Since Wicksell's pure credit economy was a closed one and there was no government sector, the equality of saving and investment means that the normal rate of interest is the macroeconomic equilibrium rate. Further, if the market interest rate equals the normal rate, there will be no change in the credit extended by banks and, therefore, no change in the money supply. For a stable amount of credit and money supply in the economy, the price level will remain unaltered. To conclude, at the market rate of interest equal to the normal one, there is equilibrium in the commodity market, Further, with stable output and money supply, the normal rate of interest will be accompanied by a stable price level.

Firms borrow to finance additions to their physical capital. The marginal productivity of capital specifies the internal rate of return to the firm's investments and was referred to by Wicksell as the *natural rate of interest*. The firm's production function has diminishing marginal productivity of capital, so that, with a constant labor force and unchanged technology, the natural rate of interest decreases as capital increases in the economy.

To see the mechanics of this model, start from an initial position of equilibrium in the economy, with a stable money supply and prices, and with the equality of the market/loan and natural rates of interest at the normal/equilibrium rate of interest. Now, suppose that while the market rate of interest is held constant by the banks, the marginal productivity of capital rises. This could occur because of technological change, discovery of new mines, a fall in the real wage rate, etc. Firms can now increase their profits by increasing their capital stock and production. To do so, they increase their investments in physical capital and finance these by increased borrowing from the banks. This causes the amount of credit and money supply in the economy to expand.

Wicksell appended to this analysis the disaggregation of production in the economy between the capital goods industries and the consumer goods industries. As the demand for investment in physical capital increases, factors of production are drawn into such industries from the consumer goods industries, so that the output of the latter falls. At the same time, the competition for labor and the other factors of production will drive up workers' incomes, leading to an increase in the demand for consumer goods, thereby pushing up prices. Consequently, the price level will rise, though with a lag behind money supply changes. Analysis based on this disaggregation of production between the capital goods industries and the consumer goods industries is not a feature of most modern macroeconomic models.

Cumulative price increases (the inflationary process)

In the above process, initiated by a reduction by the banks of the market interest rate below the natural one or by an increase in the latter above the market rate, the price rise will continue as long as the market rate of interest is below the natural rate, since the firms will then continue to finance further increases in investment through increased borrowing from

the banks. This constitutes a process of cumulative price increases. These increases can only come to an end once the banks put an end to further increases in their loans or credit to firms. A closed pure credit economy does not provide a mechanism that will compel the banks to do this.

However, in an open economy where the banking system keeps gold reserves out of which deficits in the balance of payments have to be settled, gold outflows provide a limit to the cumulative price increases. In such a context, as prices continue to increase, foreign trade deficits develop, the gold reserves of banks fall and the banks raise their loan rate of interest to the natural rate to stem the outflow of gold. This is especially so if the banks hold gold as part of their reserves and the public holds gold coins circulating as currency for some transactions. In the latter case, as prices rise, the public's demand for currency will also increase and gold will flow out of the banks' reserves to the public. Such losses of the gold reserves to the public and abroad forces banks to restrict their lending to the firms by raising their loan rate to match the natural rate. This puts an end to the cumulative credit and money supply increases and therefore to the cumulative price increases.

This cumulative process can also be initiated by banks arbitrarily lowering the market rate below the natural rate, with the resultant adjustments being similar to those specified above for an exogenous increase in the natural rate. However, Wicksell viewed the bankers as being conservative enough not to change the market rate except in response to changes in their gold holdings or an exogenous change in the normal rate. Therefore, in Wicksell's view, the cumulative price increase was usually a result of exogenous changes in the marginal productivity of capital impinging on an economy whose credit structure responds with gradual and possibly oscillatory adjustments – for example, if the banks sometimes overdo the adjustment of the market rate.

Wicksell's re-orientation of the quantity theory to modern macroeconomics

Wicksell's treatment of the pure credit economy clearly re-oriented the quantity theory in the direction of modern macroeconomic analysis. Several features of this analysis are relevant to modern macro and monetary economics. Among these is Wicksell's focus on the short-run treatment of the commodity market in terms of the equilibrium between saving and investment, a focus that was later followed and intensified in the Keynesian approach, as well as in the IS–LM modeling of short-run macroeconomics. While Wicksell claimed to be a proponent of the quantity theory of money, he shifted its focus away from exclusive attention on the monetary sector, for example, as in Pigou's version of the quantity theory, to the saving-investment process. In doing so, he led the way to the formulation of current macroeconomics, with the treatment of the commodities market at its core. This was to appear later as the IS relationship of modern macroeconomics.

Wicksell introduced into macroeconomics a fundamental aspect of the modern monetary economies: loans are made in money, not in physical capital, so that the rate of interest on loans is conceptually different from the productivity of physical capital. Even if they are equal in equilibrium, they will usually not be equal in disequilibrium. These ideas led the way to the analysis of the impact that the financial institutions and especially the central bank can have on the interest rates in the economy and on national income and employment.

Wicksell's analysis of the pure credit economy also emphasized the role of interest rates and financial institutions in the propagation of economic disturbances, since they control the market interest rate, reduction in which can set off an expansion of investment, loans and the money supply and lead to a cumulative increase in prices and nominal national income.

Further, Wicksell assumed that the banking system sets the interest rate rather than the money supply as the exogenous monetary constraint on economy. This assumption was not followed by the expositions of macroeconomic theory in either the classical or the Keynesian formulations until the end of the twentieth century, since they continued to take the money supply as their exogenously determined monetary policy variable. Since the money-demand function proved to be unstable in most developed economies after the 1970s, thereby implying the instability of the LM curve, many central banks now choose to use the interest rate as the monetary policy variable and set its level, while allowing the economy to determine the money supply as an endogenous variable for the set interest rate. This practice came to be reflected in the theories offered by the new Keynesian approach after the early 1990s. Wicksell was clearly the precursor of this type of analysis.

However, compared with the Keynesians, Wicksell, just like Fisher and Pigou, did not pay particular attention to the changes in the national output that might occur in the cumulative process. While he discussed disequilibrium and transient changes in national output during this process, he was not able to shake off the classical notion that the economy will eventually be at full employment, so that his overall discussion was usually within the context of an implicitly unchanged equilibrium level of output. Given this background, Wicksell claimed that increases in the money supply are accompanied sooner or later by proportionate price increases. Keynes's *General Theory* (1936) was to question the implicit assumption of an unchanged level of output and to allow for changes in output and unemployment following a change in aggregate demand. Merging this possibility into Wicksell's cumulative process would mean that his cumulative process would possess both output and price increases (decreases) whenever the market interest rate was below (above) the natural rate.

Hence, while Wicksell claimed nominal adherence to the traditional classical approach and the quantity theory, his theoretical macroeconomic analysis differed from theirs and was quite modern in several respects. One, in terms of this theoretical analysis in terms of saving and investment, Wicksell was a precursor of the Keynesian and modern short-run macroeconomic analysis. Two, in terms of his assumption of a pure credit economy, he presaged current developments in the payments system. Three, his assumption that the financial system sets the interest rate rather than the money supply as exogenous, he was a precursor of current central bank practices and the analysis of the new Keynesian models in the last couple of decades.

However, Wicksell's analysis did have at least several deficiencies relative to current monetary economics. One, although Wicksell did approach equilibrium through the normal interest rate which equates saving and investment, he did not present a theory of aggregate demand and also did not present the analysis of the impact of changes in it on output and employment. These were to be later addressed by Keynes. Two, Wicksell did not distinguish between real and nominal interest rates, which Fisher's equation later clarified. Three, he did not pay much attention to the analysis of the demand for money, on which Keynes made very significant contributions which provide the basis for its modern mode of treatment.

2.4 Keynes's contributions

Keynes's contributions to macroeconomics

Keynes's *The General Theory* (1936) represents a milestone in the development of macroeconomics and monetary thought. His contributions were so many and so substantial

that they led to the development of the new field of macroeconomics, which had not existed in economic thought prior to *The General Theory*. These contributions also led to a new way of looking at the performance of the economy and to an emphasis on departures from its long-run equilibrium (full employment) and the establishment of the Keynesian paradigm (see Chapter 15) in macroeconomics.

Given the very many new contributions in this book, economists have debated as to which was the most important of these contributions.[17] From a modern perspective, Keynes's emphasis on aggregate demand as a major short-run determinant of aggregate output and employment seems to have had a lasting impact on economic theory and policy. Every presentation of macroeconomic theory now includes the determination of aggregate demand and its relationship, embodied in the IS curve, to investment and fiscal policy. This contribution was based on the concept of the multiplier, which was unknown in the traditional classical period. Keynes's impact on monetary policy is reflected in central banks' manipulation of aggregate demand through either the use of the money supply or/and the interest rate, in order to maintain inflation and output at their desired levels.

Again, in terms of the modern perspective, Keynes's emphasis was on decisions on production and investment being made by firms on the basis of their expectations of future demand, and on consumption by households on the basis of their expected incomes. These decisions are usually made under uncertainty, with imperfect information on the future. Following any shifts, the reactions by firms and households to changes in demand and income prospects are often faster than by heterogeneous commodity and labor markets in adjusting prices and wages, so that the economy often produces more or less than the long-run equilibrium (full employment) output that efficient (i.e. instantly adjusting) markets will ensure. The economy is, therefore, usually likely to end up with more or less than full employment. This provides the scope for the pursuit of monetary and fiscal economies to stabilize the economy. This scope is currently reflected in the espousal of Taylor-type rules for monetary policy.

Contrary to the assumptions of the quantity theory, *The General Theory* asserted the *usual* absence of full employment in the economy. This is clearly a factual issue, which is undeniable in the context of the Great Depression of the 1930s and in many recessions. In the context of actual employment below the full-employment level, Keynes argued that output and employment depended on the aggregate demand for commodities, which, in turn, depended on the money supply, so that money was not neutral. In the context of the lengthy post-war booms in the Western economies, the contribution of high and rising aggregate demand in pushing output and employment beyond their full-employment rates is also generally recognized. The current manifestation of this recognition can be seen in the pursuit by central banks of Taylor-type rules, in which the output gap can be positive (with output above its full-employment level) or negative, with appropriate increases and decreases in interest rates expected to reduce the output gap.

Keynes, in his earlier (pre-1936) writings, had proved to be an able and innovative exponent of the quantity theory in its Cambridge school version. He had also extensively explored the effects of changes in the money stock, though still mainly within the quantity theory tradition, in the two volumes of his book *The Treatise on Money,* published in 1930. Keynes's approach to the quantity theory in the *Treatise*, as in Wicksell's writings, was in terms of saving and investment. In *The General Theory*, Keynes extended this saving-investment

17 Samuelson's (1946) obituary article on Keynes provides very valuable insights into Keynes's contributions.

approach, while abandoning the quantity theory and the traditional classical approach generally.

This chapter mainly examines Keynes's contributions on the demand for money in *The General Theory*. As a prelude to these, remember that Pigou's basic reasons for the demand for money balances were the "objects" of the provision of convenience and the provision of security. Keynes re-labeled "objects" as "motives" for holding money balances and categorized them as the transactions, precautionary and speculative motives. Of these, the transactions motive corresponded basically to the provision of convenience "object" of Pigou and the precautionary motive corresponded basically to the provision of security "object" of Pigou. Keynes was more original with respect to his speculative motive and his analysis of the demand for money balances arising from this motive.

2.4.1 Keynes's transactions demand for money

Keynes defined the transactions motive as:

> *The transactions-motive*, i.e. the need of cash for the current transaction of personal and business exchanges.
>
> (Keynes, 1936, Ch. 13, p. 170).

The transactions motive was further separated into an "income-motive" to bridge the interval between the receipt of income and its disbursement by households, and a "business-motive" to bridge the interval between payments by firms and their receipts from the sale of their products (Keynes, 1936, Ch. 15, pp. 195–6). Keynes did not present a rigorous analysis of the transactions and precautionary motives but "assumed [them] to absorb a quantity of cash which is not very sensitive to changes in the rate of interest as such … apart from its reactions on the level of income" (Keynes, 1936, p. 171). This assumption of Keynes was in fact somewhat more restrictive than that of Pigou where the demand for money, due to the objects of the "provision of convenience" and the "provision of security," was dependent upon the return on investments and the utility foregone in abstaining from consumption. Designating the *joint* transactions and precautionary demand for money balances as M^{tr} and nominal income as Y, Keynes assumed that:

$$M^{\mathrm{tr}} = M^{\mathrm{tr}}(Y) \tag{17}$$

where M^{tr} increases as Y increases.

Now consider the ratio (Y/M^{tr}), which is the velocity of circulation of transactions balances alone in the preceding equation. Here, Keynes followed the simplistic pattern of Pigou's reasoning in stating that

> There is, of course, no reason for supposing that $V(= Y/M^{\mathrm{T}})$ is constant. Its value will depend on the character of banking and industrial organization, on social habits, on the distribution or income between different classes and on the effective cost of holding idle cash. Nevertheless, if we have a short period of time in view and can safely assume no material change in any of these factors, we can treat V as nearly enough constant.
>
> (Keynes, 1936, p. 201).

This reasoning implies that Y/M^{tr} is a constant k, independent of income and interest rates, so that Keynes's transactions demand for money was:

$$M^{tr} = kY \tag{18}$$

The modern analysis of transactions demand did not follow Keynes's simplistic assumption on its constancy, but applies inventory models to it, which makes this demand a function of the interest rate. This analysis is presented in Chapter 4.

2.4.2 Keynes's precautionary demand for money

Keynes's second motive for holding money was the precautionary one, defined by him as

> the desire for security as to the future cash equivalent of a certain proportion of total resources.
>
> (Keynes, 1936, Ch. 13, p. 170).

Another definition of this motive was given later in Chapter 15 of *The General Theory* as

> To provide for contingencies requiring sudden expenditure and for unforeseen opportunities of advantageous purchases, and also to hold an asset of which the value is fixed in terms of money.
>
> (Keynes, 1936, Ch. 15, p. 196).[18]

That is, the precautionary motive arises because of the uncertainty of future incomes, as well as of consumption needs and purchases. These require holding money, an asset with a certain value, to provide for contingencies that suddenly impose payment in money. These contingencies could come from a sudden loss of income due to the loss of one's job, or a sudden increase in consumption expenditures, such as from becoming ill and requiring treatment.

Under uncertainty, the individual will form subjective expectations on the amounts required for his future payments and income receipts, and their dates, and will decide on the optimal amounts of his money balances and other assets in the light of these expectations. The further ahead are the dates of anticipated expenditures and the greater is the yield from investments, the more likely is the individual to invest his temporarily spare funds in bonds and decrease his money holdings. Conversely, an increase in the probability of requirement in the near future will lead him to increase his money holdings and decrease his bond holdings.

Although Keynes provided the rationale for the precautionary motive for holding money, he did not present a theoretical derivation of the precautionary demand for money. Rather, he merged it with the transactions demand for money. However, subsequent developments on money demand did come up with several models of the precautionary demand for money and its related buffer stock demand (see Chapter 6).

18 Keynes provided another definition of the precautionary motive on page 169 of Chapter 13. This definition differs from that in the above quotation from Chapter 15. The modern interpretation of the precautionary motive is as given in the text below the quotation.

2.4.3 Keynes's speculative money demand for an individual

Keynes's third motive for holding money was:

> 3. *The speculative-motive*, i.e. the object of securing profit from knowing better than the market what the future will bring forth.
>
> (Keynes, 1936, Ch. 13, p. 170).

Keynes had earlier explained this motive as resulting:

> from the existence of uncertainty as to the future of the rate of interest, provided that there is an organized market for dealing in debts. For different people will estimate the prospects differently and anyone who differs from the predominant opinion as expressed in market quotations may have a good reason for keeping liquid resources in order to profit, if he is right … the individual who believes that future rates of interest will be above the rates assumed by the market, has a reason for keeping liquid cash, whilst the individual who differs from the market in the other direction will have a motive for borrowing money for short periods in order to purchase debts of longer term. The market price will be fixed, at the point at which the sales of the "bears" and the purchases of the "bulls" are balanced.
>
> (Keynes, 1936, Ch. 13, pp. 169–70).

In this motive, the individual makes a choice between holding money, which does not pay interest, and bonds, which provide an uncertain return, on the basis of maximizing the return to his portfolio. With a given amount to invest in bonds or hold in money balances, he is concerned with the maturity value – equal to the capital invested plus accumulated interest – of his portfolio at the beginning of the next decision period. Assuming such a value to be uncertain, Keynes postulated a rather simple form of the expectations function: the individual anticipates a particular rate of interest to exist at the beginning of his next decision period, thereby implying a particular expected price, without dispersion,[19] for each type of bond. If these expected bond prices plus the accumulated interest are higher than the current prices, he expects a net gain from holding bonds, so that he will put all his funds in bonds rather than in money which was assumed not to pay interest and therefore to have zero net gain. If he expects a sufficiently lower price for bonds in the future than at present to yield a net loss[20] from holding bonds, he will put all his funds into money balances since there is no loss from holding these. Consequently, a particular *individual* will hold either bonds or money but not both simultaneously.

Since individuals tend to differ in their views on the future of the rate of interest, some would expect an increase in bond prices and are labeled as *bulls* in bond market parlance, choosing to increase their bond holdings, while others would expect a decrease in bond prices and are labeled as *bears*, choosing to reduce their bond holdings. Any increase in bond prices will exceed the expectations of some bulls – that is, convince them that bond prices have gone up too far and convert them into bears. A preponderance of bulls in the bond market pushes up the prices of the bonds and pushes down the rate of interest. This movement converts an

19 This simplification was subsequently abandoned in the 1950s by monetary economics in the application of portfolio selection analysis to the speculative demand for money, presented in Chapter 5 below.

20 There will be a net loss if the capital loss is greater than the interest income from holding the bond.

increasing number of bulls (who want to buy and hold bonds) into bears (who want to sell bonds and hold money), until an equilibrium price of bonds is reached where the demand for bonds just equals their supply. Therefore, the demand for speculative money balances – by bears – increases as the prices of bonds rise, or conversely, as the interest rate falls, so that the aggregate speculative demand for money is inversely related to the rate of interest.

Modern monetary and macroeconomic theory has abandoned this line of reasoning and has instead opted for an analysis based on portfolio selection, so that a better name for the money demand derived from portfolio selection analysis would be the "*portfolio demand for money*." This approach is presented in Chapter 5.

Tobin's formalization of Keynes's speculative money demand for an individual

Tobin's (1958)[21] formalization of Keynes's speculative demand analysis has become a classic and is presented in the following.

As with Keynes's analysis, Tobin assumes that there are only two assets, money and bonds, in which the individual can invest the amount of funds in his portfolio. Money is assumed to have a known yield of zero and is therefore riskless in the sense of possessing a zero standard deviation of yield. The bond is a consol, also known as a "perpetuity" in the United States, and has the characteristic that it does not have a redemption date, so that the issuer need never redeem it but may continue to make the coupon payment on it indefinitely.

In perfect capital markets, the market price of a consol will equal its present discounted value. Therefore, the price p_b of a consol which has a nominal coupon payment c per period, and is discounted at a market rate of interest x on loans, is given by:[22]

$$\begin{aligned} p_b &= \frac{c}{1+x} + \frac{c}{(1+x)^2} + \cdots \\ &= c\left(\sum_{t=1}^{\infty} \frac{1}{(1+x)^t}\right) \\ &= c\left(\frac{1}{x}\right) = \frac{c}{x} \end{aligned}$$

Therefore, the consol's value will equal its coupon rate (in perpetuity) divided by the market rate. For a given coupon value, an increase in the market interest rate will reduce the consol's price and imply a capital loss. In the special case of a bond that has the same coupon rate as the market discount rate, $c = x$, so that its market value will equal unity, i.e. $p_b = 1$.

21 Parts of the analysis of this article are presented later in Chapter 5 on the speculative demand for money.

22 The proof uses the mathematical formula that, for $x > 0$,

$$\sum_{t=1}^{\infty} \frac{1}{(1+x)^t} = \frac{1}{x}$$

Many bonds have a finite redemption date, say n. For this, the relevant formula is

$$\sum_{t=1}^{n} D^t = \sum_{t=0}^{n} D^t - 1 = \frac{1-D^{n+1}}{1-D} - 1$$

where $D = 1/(1+x)$. Since $x > 0$, $D^{n+1} \to 0$ as $t \to \infty$.

Now assume that the market interest rate is $\$R$ per year and the consol is expected to pay a coupon $\$R$ per year in perpetuity. With a coupon of R in perpetuity, the above mathematical formula implies that the consol's present value at the market interest rate R would equal R/R and be 1.

Assume for the following analysis that the coupon payment on the consol is set at R and its current price is one dollar. Further, assume that the individual expects the market rate of return on consols to be R^e for the future, with this expectation held with a probability of one and independent of the current yield R. With R treated as the coupon payment and the rate of discount expected to be R^e in perpetuity, the expected value of the consol next year will be R/R^e. Therefore, the expected capital gain or loss G on the consol will be:

$$G = R/R^e - 1$$

The expected yield $(R + G)$ from holding a consol costing \$1 is the sum of its coupon R and its capital gain G. This sum is given by:

$$R + G = R + R/R^e - 1$$

If the yield $(R + G)$ were greater than zero, the rational individual would buy only consols, since they would then have a yield greater than money, which was assumed above to have a zero yield.[23] Conversely, if the yield on consols were negative, the individual would hold only money since money would be the asset with the higher yield.

The switch from holding bonds to money occurs at $R + G = 0$. This condition can be used to derive the *critical level* R^c of the current return R such that:

$$R^c = R^c/R^e - 1 = 0$$

which implies that:

$$R^c - R^e/\left[1 + R^e\right]$$

For a given R^e, if the current interest rate R is above R^c, $(R + G) > 0$ and only consols will be bought; if it is below R^c, $(R + G) < 0$ and only money will be held. Therefore, in Figure 2.1, the individual's demand for money is the discontinuous step function (AB, CW); above R^c, the rational *individual's* whole portfolio W is held in consols and the demand for money along AB is zero; below R^c, all of W is held in money balances and the demand function is CW.

2.4.4 Keynes's overall speculative demand function

Keynes had argued that the bond market has a large number of investors who differ in their expectations such that the lower the rate of interest, the greater will be the number of investors who expect it to rise, and vice versa. Therefore, at high rates of interest, more investors will

23 Risk does not enter this choice since the individual's expectations are so firm that the subjective standard deviation of expected yields is zero. Measuring risk by the standard deviation of expected yields, the zero standard deviation means that the individual does not believe there to be any risk in holding consols.

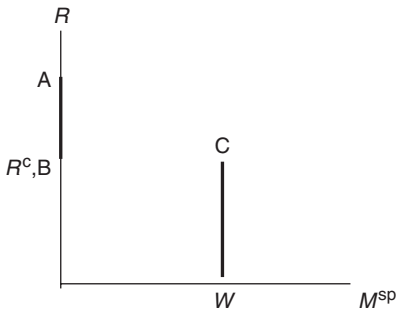


Figure 2.1

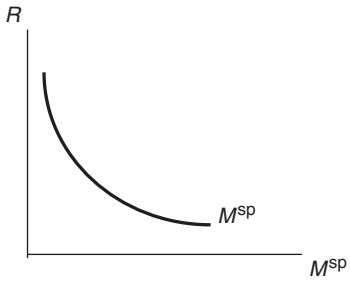


Figure 2.2

expect the rate to fall and few will hold money. At a somewhat lower rate of interest, a smaller number of the investors will expect the interest rate to fall and more of them will hold money. Hence, the aggregate demand for money will rise as the interest rate falls, and is shown as the continuous downward sloping curve M^{sp} in Figure 2.2. Therefore, Keynes's analysis implies that the speculative demand for money depends inversely upon the rate of interest, so that the speculative demand function for money can be written as:

$$M^{sp} = L(R) \tag{19}$$

where:

M^{sp} = speculative demand for money
R = market/nominal rate of interest.

Keynes called the function $L(R)$ the degree of *liquidity preference*, with L standing for liquidity.

Note that in Keynes's analysis, the individual allocates his financial wealth FW between money and bonds. Hence, in addition to the interest rate, financial wealth FW must be one of the determinants of their demand. Therefore, (19) needs to be modified to:

$$M^{sp} = L(R, FW)$$

There also exists the possibility that the economy could substitute among money, bonds and commodities as stores of value.[24] The analysis allowing this possibility would need to broaden the relevant wealth variable to total wealth and also make the speculative money demand function a function of both the return on bonds and that on commodities.[25] This extension of Keynes's analysis leads to Friedman's money demand function in the next section. For the time being, we continue with (19) for Keynes's specification of the speculative demand for money, thereby simplifying it by ignoring wealth as a determinant of the speculative demand for bonds.

2.4.5 Keynes's overall demand for money

Keynes argued that:

> Money held for each of the three purposes forms … a single pool, which the holder is under no necessity to segregate into three watertight compartments; for they need not be sharply divided even in his own mind, and the same sum can be held primarily for one purpose and secondarily for another. Thus we can – equally well, and, perhaps, better – consider the individual's aggregate demand for money in given circumstances *as a single decision, though the composite result of a number of different motives*.
>
> (Keynes, 1936, p. 195; italics added).

Hence, the aggregate demand for money, M, depends positively upon the level of income Y due to the transactions and precautionary motives and negatively upon the rate of interest R due to the speculative motive. In symbols,

$$M^{\mathrm{d}} = M^{\mathrm{tr}} + M^{\mathrm{sp}} = M(Y, R)$$

However,

> whilst the amount of cash which an individual decides to hold to satisfy the transactions-motive and the precautionary-motive is not entirely independent of what he is holding to satisfy the speculative motive, it is a safe first approximation to regard the amounts of these two sets of cash-holdings as being largely independent of one another.
>
> (Keynes, 1936, p. 199).

Hence, *as an approximation*, the demand function for money balances M^{d} is given by:

$$\begin{aligned} M^{\mathrm{d}} &= M^{\mathrm{tr}} + M^{\mathrm{sp}} \\ &= kY + L(R) = kPy + L(R) \end{aligned} \tag{20}$$

where $k > 0$ and $L(R) < 0$.

24 In the context of Keynes's analysis of bulls and bears, an expectation of inflation, with bond yields below the expected inflation rate, would mean a flight from both money and money into commodities.

25 Further, the complete analysis will also have to examine the determinants of W. In particular, as the earlier discussion on consols illustrates, financial wealth will be a function of the interest rate R.

2.4.6 Liquidity trap

Keynes argued that the speculative demand for money would become "absolute" (infinitely elastic) at that rate of interest at which the bond market participants would prefer holding money to bonds, so that they would be willing to sell rather than buy bonds at the existing bond prices. Following Keynes's reasoning, the liquidity trap occurs at the rate of interest at which a generally unanimous opinion comes into being that the rate of interest will not fall further but may rise. At this rate, there would be a general opinion that bond prices will not rise but could fall, thereby causing capital losses to bondholders, with the existing rate of interest merely compensating for the risk of such a capital loss. In such circumstances, the public would be willing to sell all its bond holdings for money balances at their existing prices, so that the monetary authorities could buy any amount of the bonds from the public and, conversely, increase the money holdings of the public by any amount, at the existing bond prices and rate of interest. Therefore, once the economy is in the liquidity trap, the monetary authorities cannot use increases in the money supply to lower the interest rate.

As against this analytical presentation of the liquidity trap, Keynes asserted that "whilst this limiting case might become practically important in future, I know of no example of it hitherto" (Keynes, 1936, p. 207). Note that this assertion was made in the midst of the most severe depression in Western history; if the liquidity trap did not exist then, it can hardly have existed in more normal periods of economic activity. Hence, in Keynes's view, while the liquidity trap is an intellectual curiosity for monetary economics, it is not of practical relevance. Consequently, contrary to some expositions or critiques of Keynesian economics in earlier decades, Keynes did not build his macroeconomic model on the assumption of the liquidity trap.

Keynes's statement on the empirical non-existence of the liquidity trap is strictly incorrect under his own analysis of the speculative demand for money. In this analysis, the liquidity trap will come into existence whenever the dominant opinion in the bond market is that the market interest rates are going to rise, not decline. Such an opinion does quite frequently come into existence in the bond markets, so that the liquidity trap is not unknown in them. Further, such an opinion can exist at any level of the rate of interest and not merely at low or even single-digit rates. Furthermore, the liquidity trap will continue to exist until the dominant market opinion changes to envision possible decreases in the rate of interest.[26] This would happen once the interest rates have adjusted to the market opinion, so that the liquidity trap would usually exist for short periods, which may not be long enough to affect investment and the macroeconomy. Therefore, while liquidity traps may often come into existence in the normal day-to-day functioning of bond markets, their existence for macroeconomics could be quite insignificant.

In contrast to Keynes's reasoning, which emphasized the possibility of a capital gain or loss on holding bonds, for the existence of a liquidity trap, is the argument that nominal interest rates close to zero do not compensate individuals for the hassle and inconvenience of holding bonds when they could forgo these by holding money balances. This argument is supported by the analysis of the transactions demand for money presented in Chapter 4, where it is shown that, at low enough interest rates relative to the brokerage costs of conversion between bonds and money, it is not profitable to hold bonds, so that only money will be

26 This, of course, is difficult to envision if the interest rate is already zero.

held; thus, in this low enough range of interest rates, the interest elasticity of money demand will be zero. In recent years, Japan is among the very few countries that have had short-term interest rates close to zero. Some empirical studies do report that the interest elasticity of money demand is much higher during Japan's low interest rate period than in other periods (see, for example, Bae *et al*., 2006).

2.4.7 Keynes's and the early Keynesians' preference for fiscal versus monetary policy

Volatility of money demand

Keynes's analysis of the speculative demand for money made it a function of the subjective expectations of the bulls and bears in the bond and stock markets. Such expectations were quite volatile in the 1930s and can be quite volatile even nowadays, as one can observe in the day-to-day volatility of the stock markets and the periodic "collapse" or sharp run-ups of prices in them. Given this volatility, Keynes asserted that the speculative demand function for money was very volatile – that is, this function shifts often. Since Keynes believed that the speculative demand for money was a significant part of the overall demand for money, the latter would also be quite volatile. This would introduce a considerable degree of instability into the aggregate demand, prices and output in the economy, and also make the pursuit of monetary policy, which could trigger changes in the investors' expectations, very risky. Keynes, therefore, was more supportive of fiscal policy than of monetary policy as the major stabilization policy in the economy. It was also the general attitude of the Keynesians until the late 1950s.

Radcliffe report: money as one liquid asset among many

The early (1940s and 1950s) Keynesians' preference for fiscal policy as against monetary policy was reinforced by the Radcliffe report[27] in Britain in 1958, which argued that money was one liquid asset among many, of which trade credit was a major part, and that the economy was "awash in liquidity," so that changes in the money supply could not be used as an effective policy tool for changing aggregate demand in the economy. Therefore, Keynes's belief in the unreliability of the effects of monetary policy (because of the instability of the money demand function) was buttressed for the 1950s Keynesians by the Radcliffe report that the money supply was only a small part of the total supply of liquidity, which was the proper determinant of aggregate demand but could not be significantly changed by monetary policy.[28]

Given the above views on the impotence of monetary policy or the unreliability of its impact, Keynesians from the 1940s to the 1960s placed their emphasis for the management of aggregate demand on fiscal policy. They advocated the active use of fiscal deficits and surpluses for ensuring the aggregate demand needed to achieve a high level of output and employment.

27 This was the report of a British parliamentary committee. It reflected the dominant Keynesian ideas on monetary policy in Britain in the late 1950s. Its conclusions on the insignificance of the impact of changes in the money supply on output were eventually not borne out empirically.

28 Monetary analysis and empirical research in subsequent decades did not support the conclusions of the Radcliffe report.

Both of the above arguments against the use of monetary policy were discarded during the 1960s when, prodded by Friedman's views and empirical findings on the money demand function, the Keynesians and the neoclassicists – and later the 1970s monetarists – came to the conclusion that monetary policy, at that time interpreted as changes in the money supply, had a strong impact on the economy. Part of this achievement was due to the contributions of Milton Friedman.

2.5 Friedman's contributions

Friedman made profound contributions to monetary and macroeconomics, especially to the role of monetary policy in the economy. He believed that monetary policy had a strong impact on output and employment, but with a long and variable lag. Among his numerous contributions was his classic article on the "restatement" of the quantity theory.

2.5.1 Friedman's "restatement" of the quantity theory of money

Milton Friedman (1956), in his article "The quantity theory of money – a restatement," sought to shift the focus of the quantity theory and bring it into closer proximity with the developments in monetary theory up to the mid-1950s. Three strands of these developments are important to note. One development was that of Keynesian macroeconomics, which placed the determination of the price level in a broad-based macroeconomic model with product, money and labor markets, and restricted the analysis of the money market to the specification of demand, supply and equilibrium in the money market. This development had argued that the price level could be affected by shifts in the aggregate demand for commodities and that changes in the money supply could affect output, and not merely prices, in an economy operating at less than full employment. The second development was Keynes's emphasis on the speculative demand for money and therefore on the role of money as a temporary store of value for the individual's wealth. The third development was the integration of the theory of the demand for money into that of goods generally by treating money as a consumer good in the consumer's utility function and as an input in the firm's production function (Patinkin, 1965).

Friedman argued that the quantity theory was merely the proposition that *money matters*, not the more specific statement that changes in it will cause proportional changes in the price level. By "money matters," Friedman meant that changes in the money supply could cause changes in nominal variables and sometimes even in real ones, such as output and employment.

Friedman restated the quantity theory to limit its main role to that of a theory of the demand for money. For consumers, the demand for real money balances was made identical to that of other consumer goods, with real balances being one of the goods in the consumer's utility function. In this role, Friedman viewed real balances as an asset, with the real values of money, stocks, bonds and physical assets being alternative forms of holding wealth and incorporated into the individual's utility function. For firms, real balances were a durable good, similar to physical capital, with both appearing as inputs in the production function. Friedman, therefore, concluded that the analysis of the demand for money was a special topic in the theory of the demand for consumer and capital goods.

Further, Friedman argued that a unit of money is not desired for its own sake but for its purchasing power over goods, so that it is a good in terms of its real and not its nominal value. This real purchasing power of money over commodities is reduced by inflation, so that the rate

of inflation is the opportunity cost of holding real balances as against holding commodities. Hence, money demand depends on the (expected) inflation rate.

Since money acts as a store of value, it is like other assets and its demand must also depend on the yield on other assets. These yields, to reflect the concern of the individual with his purchasing power, must be taken to be in their real and not their nominal value. Thus, in periods of inflation, the individual would discount the nominal yields on assets by the rate of inflation.

Friedman further argued, as in his consumption theory (the permanent income hypothesis of consumption), that the individual will allocate his lifetime wealth over commodities and over the liquidity services of real balances. This lifetime wealth (w) is the sum of the individual's human and non-human wealth, where human wealth (HW) is defined as the present discounted value of labor income while non-human wealth (NHW) consists of the individual's financial and physical assets. Since the value of these assets is known in the present, while future labor income is uncertain, the degrees of uncertainty affecting human and non-human wealth are quite different, so that their effects on the demands for commodities and money would also be different. Friedman proxied the individual's degree of uncertainty of wealth by the ratio of his human to non-human wealth.

Therefore, according to Friedman, the main determinants of the individual's demand for real balances were the *real* yields on other assets (bonds, equities and physical assets), the rate of inflation, real wealth and the ratio of human to non-human wealth. Writing this demand function in symbols,

$$m^{\mathrm{d}} = M^{\mathrm{d}}/P = m^{\mathrm{d}}(r_1, \ldots, r_n, \pi, w, HW/NHW) \tag{21}$$

where:

m^{d} = demand for money balances in real terms
M^{d} = demand for money balances in nominal terms
P = price level
r_i = yield in real terms on the ith asset
π = rate of inflation
w = wealth in real terms

HW/NHW ratio of human to non-human wealth.

Permanent income as the scale determinant of money demand

Since data on human and total wealth was not available, Friedman proxied total wealth by permanent income y^{p}. At the theoretical level, the relationship between these variables is specified by:

$$y^{\mathrm{p}} = rw \tag{22}$$

where r is the expected average real interest rate over the future. Permanent income y^{p} can be interpreted as the average expected real income over the future. In line with Friedman's work on the consumption function, Friedman employed adaptive expectations – which use a geometric lag of past incomes – to estimate y^{p}, rather than rational expectations. These procedures will be covered in Chapter 8.

Since the demand function is derived from the consumer's utility function, which represents the individual's tastes, shifts in these tastes will shift the demand function. Friedman sought

to take account of such shifts by incorporating a variable u for "tastes/preferences" in the demand function. Substituting y^{p} for w, taking r to be proxied by the various interest rates and adding the new variable u for tastes/preferences, in the manner of Friedman's article, the demand function for real balances becomes:

$$m^{\mathrm{d}} = M^{\mathrm{d}}/P = m^{\mathrm{d}}(r_1, \ \ldots, r_{\mathrm{n}}, \pi, y^{\mathrm{p}}, HW/NHW, u) \tag{23}$$

Note that this demand for money is essentially derived from the notion of money as an asset – that is, a store of value – and that permanent income appears in it as a proxy for wealth.

Friedman on the velocity of money

Since the velocity of circulation V equals Y/M, and M in equilibrium equals M^{d}, we have:

$$V = \frac{y}{m^{\mathrm{d}}(r_1, \ldots, r_n, \pi, y^{\mathrm{p}}, HW/NHW, u)} \tag{24}$$

where both the numerator and the denominator on the right-hand side of the equation are real variables, so that their ratio is also a real variable. The preceding equation implies that, for Friedman, velocity was not a constant but a real variable, which depended upon the real yields on alternative assets and other variables. Except for the introduction of permanent income instead of current income as a determinant on the right side, (24) was consistent with the Keynesian tradition. The essential difference between Friedman and Keynes was on the stability of the velocity function: Friedman asserted that velocity was a function of a few variables and the velocity function was stable, whereas, for Keynes, the velocity function possessed, by virtue of the volatile nature of the subjective probabilities on bond returns, the potential for being unstable and its shifts unpredictable.

Friedman on the money supply

On the money supply, Friedman asserted that the supply function of money was independent of the money demand function. Further, some of the important determinants of the former, including political and psychological factors, were not in the latter. Hence, the money demand and supply functions were separate and could be identified in the data.

Friedman, like Keynes, assumed that the central bank determines the money supply, so that it could be treated as an exogenous variable for the macroeconomic analysis of the macroeconomy. This is, of course, a practical question. Its validity depends on central bank behavior. By the mid-1990s, many central banks were using the interest rate as their primary monetary policy instrument, while leaving the money supply to be determined endogenously by the economy at the set interest rate. While the exogeneity of the money supply was an unquestioned mainstay of short-run macroeconomic models and of the IS–LM analyses until the mid-1990s, the new Keynesian models which have emerged since the mid-1990s tend to assume that the central bank sets the interest rate, so that the money supply becomes endogenous in these models.

2.5.2 Friedman on inflation, neutrality of money and monetary policy

On the basis of his empirical studies, Friedman asserted that inflation is always and everywhere a monetary phenomenon. This assertion has become quite famous. While it does

not accurately explain the determination of low inflation rates (i.e. in the low single digits), it does explain quite well persistently high inflation rates over long inflationary periods. The attribution of persistently high inflation rates to high money supply growth rates has already been explained in the earlier presentation of the quantity equation.

Friedman held that money was neutral in the long run. But, for the short term, he was strongly of the view that money was not neutral and, in fact, offered very significant and convincing economic evidence from the history of the United States that it was not so (Friedman and Schwartz, 1963, esp. pp. 407–19, 712–14, 739–40; Friedman, 1958). He also distinguished between anticipated and unanticipated changes in inflation rate and argued that the initial effects of a higher unanticipated inflation rate last for about two to five years, after which the initial effects start to reverse, so that the effects of unanticipated money supply and inflation increases on output, employment and real interest rates could last ten years (Friedman, 1968). To Friedman, changes in the money supply had a strong impact on output and unemployment,[29] and major depressions and recessions were often associated with severe monetary contractions.[30] Conversely, for the USA, major inflations were usually associated with wars,[31] during which the large fiscal deficits were financed by increases in the money supply.

However, Friedman maintained and showed that the timing of the impact of money supply changes on output was unpredictable and the lags involved were long and variable (Friedman, 1958). He concluded that major instability in the United States has been produced or, at the very least, greatly intensified by monetary instability. Consequently, he maintained that discretionary monetary policy could have unpredictable results and should not be followed. He claimed that:

> The first and most important lesson that history teaches us … is that monetary policy can prevent money from itself being a major source of economic fluctuations … [It] should avoid sharp swings in policy. In the past, monetary authorities have on occasion moved in the wrong direction. … More frequently, they have moved in the right direction, albeit often too late, but have erred by moving too far. … My own prescription … is that [it adopt] a steady rate of growth in a specified monetary total. … The precise rate of growth, like the precise monetary total, is less important than the adoption of some stated and known rate.
>
> (Friedman, 1968, pp. 12–16).

2.5.3 Friedman versus Keynes on money demand

Friedman's main concern in deriving his demand function was with money as a real asset held as an alternative to other forms of holding wealth, whereas Keynes's analysis was for the demand for nominal money balances. Friedman's analysis also implied that money demand depends on wealth or permanent income, rather than on current income as in Keynes's analysis. However, Friedman believed that the demand for money does not in practice become

29 Friedman shared these views with most pre-Keynesian (traditional classical) economists.

30 In the USA, during the Great Depression years 1929 to 1933, the money supply decreased by more than one-fourth, due to increases in the ratios of the public's currency holdings and of bank reserves to money supply, as well as bank failures.

31 For the USA, this was so for the Civil War, World Wars I and II, Korean War and Vietnam War.

infinitely elastic, thereby agreeing with Keynes on the absence of the liquidity trap in practice. Further, Friedman believed that the money demand function was stable, whereas Keynes had adduced the subjective nature of probabilities in the absence of complete information on the future returns on bonds to derive the volatility of the speculative and overall money demand. On this point, Friedman's own and others' empirical findings for the 1950s and 1960s data supported Friedman over Keynes on the stability of the money-demand function (see Chapter 9).

Friedman further asserted that the money-demand and velocity functions were even more stable than the consumption function.[32] Till the late 1960s, the stability of the latter was the linchpin of the Keynesian analysis in its enthusiastic support for fiscal policy over monetary policy. Friedman's assertion meant that monetary policy would, at least, also have a strong impact on the economy. The success of Friedman's agenda was such that by the early 1960s the Keynesians had accepted monetary policy as having a strong and fairly reliable impact on aggregate demand, so that a synthesis – known as the neoclassical-Keynesian synthesis – emerged in the 1960s. This synthesis was reflected in the common usage of the IS–LM model for the macroeconomic analysis of the impact of monetary policy on aggregate demand. The divisions among these schools were henceforth confined to questions of the further impact of aggregate demand changes on output and unemployment.

At a general level, Friedman's money-demand analysis was not an elaboration or restatement of the quantity theory, despite Friedman's claim for it, and could more appropriately, as Patinkin (1969) pointed out, have been labeled a statement of the Keynesian money demand function or of the portfolio approach – as in Tobin (1958) – to money demand topical in the 1950s.[33]

Friedman was essentially a Keynesian in his macroeconomic theory and on his theory of money demand, but he was a conservative on the pursuit of monetary policy (Patinkin, 1981). On macroeconomics, his theoretical and empirical contributions showed that changes in the money supply could have strong effects on both nominal and real output. On monetary policy, Friedman advocated that an active monetary policy should not be pursued. Part of this advocacy was based on his roots in political conservatism and partly on his empirical finding that money supply changes impact on the economy with a long and variable lag. Friedman's macroeconomic theory and policy recommendations are presented in Chapter 14.

2.6 Impact of money supply changes on output and employment

The standard short-run macroeconomic models establish the importance of the money supply in determining nominal national income. As we have seen in this chapter, the dominant theory on this subject in the nineteenth and early twentieth centuries was the quantity theory. It had implicit in its acceptance the classical theories on the determination of output and interest rates, both of which were outside the influence of the demand and supply of money in long-run equilibrium. However, as Hume and other economists in the eighteenth and nineteenth centuries had argued, the money supply did significantly affect output and other real variables

32 Friedman sought to establish this in his later publications, jointly with David Meiselman.

33 Patinkin, another economist at the University of Chicago in the 1950s, presented a clearer and more accurate representation of the Chicago tradition and argued that Friedman was closer to the Keynesian tradition than the quantity theory tradition. He points out the strong influence of Keynesian monetary ideas on the Chicago School during the 1940s and 1950s, especially in terms of the emphasis on the portfolio demand for money, and on Friedman's own analysis.

in the disequilibrium process through the greater availability of funds for consumption and investment.

The impact of Keynes's *General Theory* and the Great Depression led to the recognition that, in practice for a given country, output may not always be at its full-employment level. This meant that one of the critical background assumptions of the quantity theory and traditional classical economics had to be abandoned. This was the assumption that labor markets work in such a manner as to ensure *continuous* full employment of resources and hence ensure that output is *always* at its full-employment level. Keynesian analysis showed that in the presence of less than full-employment, real output and unemployment can be influenced by the policy makers through money supply changes. As pointed out earlier in this chapter, the theoretical contributions of Milton Friedman also subscribed to this proposition and his empirical studies with Anna Schwartz confirmed this possibility. This proposition became part of the theoretical Keynesian-neoclassical synthesis of the 1960s.

The possible existence of a not-so-transitory disequilibrium or equilibrium state with less than full employment, as well as the non-neutrality of systematic monetary policy, were rejected by the modern classical school of macroeconomics in the 1970s. However, this conclusion has been challenged in the last two decades by the resurgent new Keynesians. The current consensus on these issues seems to be that:

- Monetary policy (i.e. changes in money supply or interest rates) is neutral in the long run but not in the short run.
- Empirically, monetary policy impacts output and employment with a lag and usually does so earlier than its impact on prices and inflation.
- As a corollary, monetary policy often has much of its impact on output and employment without first causing a change in market prices.

Transmission mechanisms for the impact of monetary policy on output: the heritage

The mechanisms by which increases in the money supply affect nominal national income have historically been a matter of considerable dispute. David Hume (1752) had specified this mechanism as being of a dual nature. He started with the supposition that there is a sudden increase in everyone's money holdings and analyzed the transmission channel of its effects on national income and expenditures in the following terms:

> The prodigal landlord dissipates it as fast as he receives it and the beggarly peasant has no means, nor view, nor ambition of obtaining above a bare livelihood. The overplus of borrowers above that of lenders continuing still the same, there will follow no reduction of interest.
>
> Hume (*Of Interest*, 1752).

However, if the increase in the money supply falls into a few hands:

> [and is] gathered into large sums, which seek a secure revenue either by the purchase of land or by interest … the increase of lenders above the borrowers sinks the interest. … But after this new mass of gold and silver has been digested and has circulated through the entire state … [the rate of interest will return to its former level].
>
> (Hume, *Of Interest*, 1752).

Hume thus emphasized two channels of influence of increases in the money supply. One of these was through increased spending on commodities, mainly by those whose consumption absorbs virtually all their incomes. This channel is now known as the direct transmission channel. The other, indirect transmission channel operated through the increased availability of loanable funds. The second channel operated mainly if the initial increase in the money supply ended up in lump sums in the hands of lenders, whose modern counterpart is mainly financial institutions. The relative strength of each channel depended upon the structure of the economy and the diffusion of the new money balances.

Irving Fisher mainly emphasized the direct transmission channel, as in:

> [Suppose that an individual's money holdings are doubled.] Prices being unchanged, he now has double the amount of money and deposits which his convenience had taught him to keep on hand. He will then try to get rid of the surplus money and deposits by buying goods. … Everybody in the community will want to exchange this relatively useful extra money for goods, and the desire to do so must surely drive up the price of goods.
>
> (Fisher, 1911).

2.6.1 Direct transmission channel

This transmission channel, whereby increases in the money supply cause undesired money balances which are then directly spent on commodities, is called the *direct transmission channel* and is associated with the followers of the quantity theory. Among these were Milton Friedman and the monetarist school of the 1970s. However, the modern classical school has not followed the 1970s monetarists in this respect and has stayed with the indirect transmission mechanism in its models. Part of the reason for this lies in the nature of the modern economy in which changes in the money supply are not disbursed directly to households but initially enter the financial markets, often through open-market operations.

2.6.2 Indirect transmission channel

The Keynesian tradition and the IS–LM macroeconomic models ignore the direct transmission channel. The closed-economy versions of these models assume that total expenditures are composed of consumption, investment and government expenditures. In these models, consumption depends upon real income, investment depends upon the rate of interest and government expenditures are exogenously determined. None of these major components of total expenditures depend *directly* upon the availability of money, so that increases in the latter are not directly spent on any of those components. Money supply increases affect the economy by lowering interest rates, which increase investment, which in turn pushes up nominal national income through the multiplier in the commodity markets. This mode of transmission of money supply increases, through interest rates and investment, to national expenditure and income increases is known as the *indirect transmission channel*.

Central banks now rely on open-market operations for changing the money supply, which change the interest rate, or/and set the interest rate for the economy. The neoclassical and modern classical schools also do so for their analyses. Therefore, macroeconomic policy and models nowadays incorporate the indirect transmission channel but not the direct transmission one.

2.6.3 Imperfections in financial markets and the lending/credit channel

In perfect capital markets, with full information on borrowers, lenders need only to rely on the interest rate charged on loans, since this includes all the available information on the risks involved in making the loan to the borrower and compensation for that risk. However, the lack of full information leads lenders to limit their loans to a particular borrower. The transmission channel associated with imperfections in financial markets is known as the lending/credit channel.

Some of the total borrowing in the economy occurs in the form of loans by banks to their customers, whether firms or households, by suppliers to the buying firms in the form of trade/business credit, and from households to small firms. In these loans, the interest rate is usually only one of the elements of the loan. Another is the lender's belief in the credit-worthiness of the borrower, which determines the riskiness of the bonds issued by the borrower, so that lenders can reduce their risks by rationing the amount lent to any given borrower or by making the loans only to certain categories of borrowers. In view of this aspect of direct loans, some economists distinguish between the flows of funds through the bond (including the equity) market and those through loans/credit, thereby creating a distinction between bonds and loans/credit as distinctive assets, with less than perfect substitution between them (Bernanke and Blinder, 1988; Kashyap and Stein, 1993, 1997). Given the special aspects of direct loans, loans usually offer a higher return than bonds to the lender while enabling borrowers, who may not have access to the bond market, to obtain funds.

In models that separate credit from bonds, the aggregate demand and output of commodities responds to changes in both the money supply and the loan supply. For the financially developed economy, while some economists believe the lending/credit channel to be a significant and distinct factor in the effects of monetary policy (Hubbard, 1995; Kashyap and Stein, 1993, 1997), most of the profession has tended to be skeptical about its relative significance for financially developed economies (Miron *et al*., 1994).

The analysis of the credit market is given in Chapter 16. Chapters 1 and 16 also provide information on the credit crisis in the market for asset-backed corporate bonds in the USA in 2007. This crisis provides a striking example of the impact of credit and money markets on the real sectors of the economy, and the limited ability of the monetary authorities to offset a credit crunch.

2.6.4 Review of the transmission channels of monetary effects in the open economy

We have so far discussed the direct channel and the indirect and credit channels operating through the bond and credit interest rates. An additional channel occurs through the impact of changes in expectations on aggregate demand and the inflation rate, with the role of the latter due to the Fisher equation on interest rates, which asserts that the nominal rates incorporate the expected rate of inflation. For the modern open economy, an additional channel operates through exchange rate changes.

Therefore, the impact of monetary policy on aggregate demand and output occurs in various ways, the most important of which are:

- direct transmission, due to the spending of excess money balances;
- indirect transmission through interest rates on bonds;
- indirect transmission through the amount of loans/credit and the interest rate on them;

- indirect transmission through the expected rate of inflation;
- indirect transmission through the exchange rate.

The inflation-expectations route takes account of the impact of monetary policy on the expectations of future inflation by forward-looking economic agents and is often used to differentiate between the impact of anticipated and unanticipated monetary policy on output. The exchange-rate channel operates in the open economy because monetary policy changes the domestic interest rates (which affect net capital flows to the domestic economy) and aggregate demand (which changes the exchange rate, prices and net exports). Chapter 13 discusses these effects. The loan/credit channel is discussed in Chapter 16.

Note that each channel introduces its own pattern of lags in the impact of monetary policy on aggregate demand and output. In addition, different countries will have a different relative role of the various channels, different lag structures and a different overall impact of monetary policy, at least in the short run.

2.6.5 Relative importance of the various channels in financially less-developed economies

Some countries, mainly among the LDCs, have both a large informal financial sector and large, legally unaccounted funds. The latter are known as "black money." Both of these enhance the significance of the lending and direct transmission channels. Black money cannot be deposited in formal financial institutions, where it could become loanable funds, nor can it be used to directly buy publicly traded bonds. Its common use is to buy commodities, for which payment can be made wholly or partly in money. If loans are made by the holders of black money, they often depend primarily on personal knowledge, including that of trustworthiness, of the borrower, and only secondarily on the interest rate. These factors reduce the significance of the indirect channel (based on market interest rates) relative to the direct and lending channels since some of the increases in the money supply are likely to percolate to the informal sector and some to end up as black money. Therefore, while the indirect transmission (through the bond rate of interest) channel is likely to be the more important one in the developed economies, the direct transmission and lending channels can be quite important in any given LDC.

Conclusions

The quantity theory tradition in the eighteenth to the early twentieth centuries was the result of the thinking of many great economists over more than a century. There was a great deal of variation around the central theme of the quantity theory that changes in the money supply cause proportional changes in prices and nominal national income in equilibrium. While Friedman was right that the quantity theory was a living tradition and not a rigid doctrine throughout the traditional classical period (until Keynes), the profession has not followed him in viewing his monetary theory as an aspect of the quantity theory. Therefore, the quantity theory should now be viewed as a historical but not a modern doctrine; it is not now a living tradition.

Friedman had also proposed that the quantity theory was a theory of the demand for money, rather than a theory of the price level. From the perspective of the history of economic ideas, the profession has refused to accept this re-orientation of the quantity theory. The consensus remains that it was a theory of price determination, i.e. that in equilibrium, changes in the

money supply cause proportionate increases in the price level but do not affect output and employment, rather than merely being a theory of the demand for money.

In many ways, Friedman's (1956) statement of the demand for money function was not a statement of the historical doctrine of the quantity theory[34] but rather a topical synopsis of the ideas and developments in monetary theory up to the 1950s. In particular, Friedman followed the Keynesians in their emphasis on money as an asset acting as a temporary store of value – and therefore, one asset among many. This represented an undue emphasis on the store-of-value component of Keynes's analysis, to the unwarranted downplaying of the transactions role of money.

The distinctive aspects of Friedman's 1956 article on the quantity theory were his assertions that the money-demand function and the velocity function are stable. This was a radical departure from Keynes and the Keynesian opinions of the 1950s, which viewed money as merely one liquid asset among many and considered the money-demand function to be highly volatile because of the psychological basis of the probabilities of the return on bonds. Friedman's success in this tournament was sufficient to lead the profession – including the Keynesians – to accept by the 1960s that money matters and so does monetary policy.

Knut Wicksell's contributions were seminal in making the saving-investment process central to the analysis of the impact of money supply changes on the economy. Unlike some of the traditional versions of the quantity theory – which had kept the monetary analysis separate or used a dichotomy between the real and the monetary sectors – Wicksell integrated the monetary analysis with that of the commodity and bond markets. This approach led the way to the later Keynesian and neoclassical macroeconomic models and also foreshadowed those macroeconomic analyses of the last two decades that are based on the interest rate as the operating target of monetary policy.

Keynes had left the quantity theory's representation of the transactions demand for money essentially unchallenged. As an asset not itself used for consumption or production but held for financing transactions, money balances held by an individual have all the characteristics of inventories of goods held for production or sale. Baumol and Tobin (see Chapter 4) later formulated the transactions demand for money balances by applying to it the basic theory of the demand for inventories. This analysis is the subject of Chapter 4.

Keynes's motives for holding money and their analyses were a continuation and elaboration of the ideas of the Cambridge school, with the highly distinctive addition of the speculative demand for money. The fervent in monetary theory raised by Keynes's emphasis on the speculative demand for money was, as seen above, essentially a formalization of some aspects of the store-of-value function of money, that is, of money as an asset among many forms of holding wealth. Friedman (1956) presented a compact statement of the further developments of this approach. Formal and rigorous analysis of risk-taking in the demand for money and other financial assets was subsequently developed by the application of the expected utility theory of portfolio selection. This analysis is presented in Chapter 5. The precautionary demand and the buffer stock analyses of money demand are presented in Chapter 6.

The ferment in the theoretical analysis of the demand for money, initiated by Keynes, ran out of steam by the 1980s, and innovations in its empirical estimation did so by the 1990s. Relatively few contributions on this topic have appeared since the 1980s.

34 Patinkin (1969) cogently argued that Friedman (1956) "provided us with a most elegant and sophisticated statement of modern Keynesian monetary theory – misleadingly entitled 'The quantity theory – A Restatement'." (Patinkin, 1969, p. 61).

Summary of critical conclusions

- The quantity theory consisted of several approaches in its evolutionary history. They asserted that, in long-run equilibrium, a change in the money supply would cause a proportionate change in the price level but would not affect output and unemployment.
- In disequilibrium, the quantity theory allowed changes in the money supply to affect output and employment.
- Wicksell shifted the transmission mechanism (from money to aggregate demand) from the direct transmission mechanism to the indirect one. Further, he envisaged the banking system as setting the interest rate, thereby making the money supply endogenous.
- Keynes expanded the reasons for holding money to encompass the transactions motive, the precautionary motive and the speculative motive.
- Friedman, although ostensibly claiming to provide a "restatement" of the quantity theory, in fact provided an integrated version of the neoclassical and the Keynesian ideas on the demand for money. However, his replacement of current income by permanent income as the scale determinant of money demand belonged in neither the quantity theory nor the Keynesian traditions.
- Keynes and the Keynesians integrated the analysis of the money market and the price level into the general macroeconomic model, rather than leaving it as an appendage to the analysis of the commodity markets. They also introduced bonds as an alternative asset to money in the demand for money and made the bond market a component of macroeconomic analysis.
- There are several potential transmission mechanisms through which changes in the money supply impact on aggregate demand. Their basic classification is into the direct transmission mechanism and the indirect one.
- Whether the lending channel is distinct from the indirect one through interest rates and whether it is significant for the modern financially developed economies is still in dispute.
- While the money supply used to be the primary operating target of monetary policy, this target is now the interest rate, so that the money supply becomes endogenous with respect to the set interest rate.

2.7 Review and discussion questions

1. Discuss the statements:
   "The quantity theory and the quantity equation are one and the same in the sense that each implies the other."
   "The quantity theory assumes the constancy of velocity."
2. Compare the approaches of Fisher's transactions and Pigou's cash balances to the quantity theory. Are there any similarities between them? If so, in which respects? Or should they be treated as different approaches altogether?
3. Given Pigou's elucidation of his two "provisions" for holding money, was Keynes's exposition of his three "motives" a revolutionary change or merely an extension of the money demand analysis to an economy in which the bond and stock markets were becoming increasingly visible and significant for the macro economy? Discuss.
4. Compare the contributions of Pigou, Keynes and Friedman on the interest elasticity of the demand for money.
5. Discuss the following statement: Wicksell's analysis of the pure credit economy belongs in the Keynesian rather than the quantity theory tradition, so that Wicksell's analysis should be taken as the precursor of Keynesianism in monetary economics.

6. Discuss the following statement: Friedman's analysis of the demand for money belongs in the Keynesian rather than the quantity theory tradition, so that his analysis should be taken as a statement of, or slight modification to, Keynesian ideas in monetary economics.
7. Can overall money demand be legitimately separated into three additive components according to Keynes's motives for holding money? If not, what is the justification for doing so?
8. For Keynes, the speculative component of money demand was volatile. This made the demand for money and the money multiplier volatile, so that monetary policy became an unreliable tool for stabilization. What were Keynes's reasons for his assertion on volatility? Do you think such volatility exists in the modern economy? Has it increased or decreased over time?
9. For Friedman, the money-demand function was highly stable. This made the money-income multiplier highly stable, so that changes in the money supply had a strong impact on nominal national income. What were Friedman's reasons for his assertion on the stability of money demand?

   Has the money demand function in recent years been stable in the sense of not possessing the type of volatility asserted by Keynes? Discuss.

   In many economies in recent decades, the money-demand function has shifted over time due to financial innovations. Have these shifts invalidated Friedman's assertion, or is this instability of a different kind from what Friedman and Keynes had in mind?
10. What were the similarities and differences between Keynes's and Friedman's demand functions for money? In which tradition (Keynesian, quantity theory or traditional classical) did Friedman's analysis belong?

    Discuss Friedman's views on velocity and present his velocity function.
11. Discuss the following statement: Friedman's critique of Keynesian liquidity preference theory, and especially of the Keynesian speculative motive, is more concerned with the stability rather than with the interest elasticity of money demand.
12. On what does the demand for money depend: current income, wealth or permanent income? Or does it directly depend upon neither of them but on the consumption expenditures of households and the output of firms? If so, why is money demand usually specified as a function of income?
13. What were the views of Keynes and Friedman on the exogeneity or endogeneity of the money supply? What justifies their views?

    What were the views of Wicksell on the exogeneity or endogeneity of the money supply? What justifies his views?

References

Bae, Y., Kakkar, V. and Ogaki, M. "Money demand in Japan and non-linear cointegration." *Journal of Money, Credit and Banking*, 38, 2006, pp. 1659–67.

Bernanke, B.S., and Blinder, A.S. "Credit, money and aggregate demand." *American Economic Review,* 78, 1988, pp. 415–39.

Fisher, I. *The Purchasing Power of Money*. New York: Macmillan, 1911, Chs 1–4, 8.

Friedman, M. "The quantity theory of money – a restatement." In M. Friedman, ed., *Studies in the Quantity Theory of Money*. Chicago: Chicago University Press, 1956, pp. 3–21.

Friedman, M. "The role of monetary policy." *American Economic Review*, 58, 1968, pp. 1–17.

Friedman, M. "The supply of money and changes in prices and output" (1958). Reprinted in M. Friedman, *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine, 1969.

Friedman, M., and Schwartz, A. *The Monetary History of the United States, 1867–1960*. Princeton, NJ: Princeton University Press, 1963.

Hubbard, R.G. "Is there a credit channel for monetary policy?" *Federal Reserve Bank of St Louis Review*, 77, 1995, pp. 63–77.

Hume, D. *Of Money* (1752). Reprinted in *The Philosophical Works of David Hume*. 4 volumes. Boston: Little, Brown and Co., 1854.

Hume, D. *Of Interest* (1752). Reprinted in *The Philosophical Works of David Hume*. 4 volumes. Boston: Little, Brown and Co., 1854.

Kashyap, A.K., and Stein, J.C. "Monetary policy and bank lending." *NBER Working Paper No*. 4317, 1993.

Kashyap, A.K., and Stein, J.C. "The role of banks in monetary policy: A survey with implications for the European monetary union." *FRB of Chicago Economic Perspectives,* 21(5), 1997.

Keynes, J.M. *Treatise on Money.* 2 volumes. New York: Harcourt, Brace, 1930.

Keynes, J.M. *The General Theory of Employment, Interest and Money*. London and New York, 1936, Chs 13, 15.

Miron, J.A., Romer, C.D. and Weil, D.N. "Historical perspectives on the monetary transmission mechanism." In N.G. Mankiw, ed., *Monetary Economics*. Chicago: University of Chicago Press, 1994.

Patinkin, D. *Money, Interest and Prices*. 2nd edn. New York: Harper & Row, 1965, Chs 2, 5–8.

Patinkin, D. "The Chicago tradition, the quantity theory, and Friedman." *Journal of Money, Credit and Banking*, 1, 1969, pp. 46–70.

Patinkin, D. *Essays on and in the Chicago Tradition*. Durham, NC: Duke University Press, 1981.

Pigou, A.C. "The value of money." *Quarterly Journal of Economics*, 32, 1917, pp. 38–65.

Samuelson, P. "Lord Keynes and the General Theory." *Econometrica*, 14, 1946, pp. 187–200.

Tobin, J. "Liquidity preference as behavior towards risk." *Review of Economic Studies*, 25, 1958, pp. 65–86.

Wicksell, K. "The influence of the rate of interest on prices." *Economic Journal*, 17, 1907, pp. 213–20.

Part II

Money in the economy

3 Money in the economy

General equilibrium analysis

This is a core analytical chapter from a microeconomic perspective on money in the economy. It treats real balances as a good like other goods such as commodities and labor in the economy and derives their demand in the overall context of the demands and supplies for all goods in the economy. It uses these in a Walrasian general equilibrium model to determine the relative and absolute prices of goods, and examines their properties. In particular, it addresses rigorously the controversial and important questions of the neutrality and super-neutrality of money.

While the analysis of this chapter is based in microeconomics, its conclusions apply to both the microeconomics and macroeconomics of money. Therefore, this chapter can be covered immediately after Chapter 2 as a continuation of the heritage of monetary economics to the Walrasian model. Alternatively, it can be covered after Chapter 12 and thereby be a precursor to the macroeconomics Chapters 13 to 17 on money in the macroeconomy.

The analysis of this chapter is fundamental to a rigorous consideration of the foundations of monetary theory.

Key concepts introduced in this chapter

- Definition of a good in economics
- Money as a good
- The demand for real balances
- Numeraire
- User cost of money
- Money in the utility function (MIUF)
- Money in the production function (MIPF)
- Relative versus absolute prices
- Homogeneity of degree zero
- Neutrality of money
- Super-neutrality of money
- Dichotomy between the real and monetary sectors
- The real balance effect

This chapter considers real balances as a "good" in economics and presents its analysis. In conformity with preference-based analysis in economics, it defines a good as anything

of which more is desired to less. It then lists the stylized facts about money in a monetary economy, which also has labor, commodities and bonds among other goods. It then presents three models to derive the demand for money. If a model's implications do not meet the stylized facts, it is rejected as inappropriate for monetary economies. This is so for a model commonly used in macroeconomics.

This chapter then derives the demand for money as an element of the Walrasian (general equilibrium) model, which forms the foundation of the microeconomic analysis of the markets of the economy and the determination of individual prices. It also forms the basis of the modern classical and neoclassical macroeconomic models, and the standard against which Keynesian macroeconomics lays out its differences. This chapter therefore serves as the prelude to Chapters 4 to 10 on the microeconomic aspects of monetary analysis – that is, the demand and supply of money. It also serves as a prelude to Chapters 13 to 17 on the macroeconomic aspects of monetary economics – that is, money and monetary policy in macroeconomic models. The reader more interested in these macroeconomic aspects can, therefore, proceed after this chapter directly to Chapters 13 to 18.

3.1 Money and other goods in the economy

Definition of a "good"

To consider whether money is a good or not, we need a definition of "goods." From the analysis of the behavior of individuals or households, we define a *good* as something of which an individual desires more rather than less, or less rather than more, *ceteris paribus*. A particular good may or may not be marketed; thus silence may be a good in the midst of overwhelming noise and yet may not be marketed.[1] From the point of view of the relevance to a market economy, only those goods that are marketed at some price or other need to be considered. Further, note that economic analysis does not ask why more of a good is desired to less of it. Therefore, it does not need to consider whether the good is in some sense beneficial or injurious for the individual, or whether there is something innate to the individual as a biological entity or something in the social or physical environment, or any other factor, which affects the individual's desire for its acquisition. To take some odd examples, diamonds, cigarettes, drugs, labor time spent in a criminal activity, guns and bombs, etc., are all treated as goods (or "bads") in microeconomic analysis. So is money, though it is not "directly consumed" and even though its components (such as the currency of the particular country and the demand deposits in it) only constitute money by virtue of the social and economic environment that make them acceptable as a medium of payments. Note that this is also so for diamonds, as for many other commodities, whose demand arises not because they or their services are "directly used in consumption or production" but because of the social and economic environment which creates utility for them or their services. The desire of an individual to hold diamonds or real balances constitutes adequate reason for treating them as goods in his utility function. The fact that money can only be held and used at a cost only adds

1 It is, however, marketed in some cases, as in the case of "soundproof apartments" commanding higher rents than other apartments.

confirmation to the treatment of money as a good for individuals, but is strictly not necessary to this treatment.

From the point of view of a firm, an input (which is a type of good) is anything of which more rather than less increases (or decreases) its production. Economic theory does not ask why it does so and, therefore, does not consider whether a good "directly" enters production or whether more or less of it increases production by virtue of the environment in which the firm functions. The desire of firms to hold real balances constitutes an adequate reason for treating money as an input to their production, so that it constitutes a good for them.

Money and other goods in macroeconomic analysis

For macroeconomic modeling, *goods* are subdivided into the categories of *commodities* or *products*, *labor* or its converse as *leisure, money* and *bonds*, where the term "bonds" is defined to encompass all *non-monetary financial assets*.[2] Compared with other goods, *money* is the most liquid good and serves as the medium of payments. This chapter assumes that commodities, labor and bonds are relatively illiquid goods and cannot be used directly for exchanges against commodities.

One general trend of thought throughout the nineteenth century, and increasingly since the 1930s, argued that the demand for money should be analyzed as that of a choice of one good among many. This approach claimed that the analytical framework for determining the demand for real balances to be held by an individual or firm is the same as that for determining the demand for commodities in general, and that this framework is that of utility maximization for the individual or household and profit maximization for the firm. This approach is at present the dominant one in monetary theory, and Friedman's (1956) version of it was presented in Chapter 2. Such an approach can be formulated in terms of a timeless analysis, a one-period one or an intertemporal one.

Different approaches to deriving the demand for money

There are three main approaches to deriving the demand for money and its role in the economy. These are:

1. Money yields utility and can therefore be incorporated into the utility function. Similarly, money can be incorporated into the production function. Alternatively, while money is not directly a component of the utility and production functions, it saves labor time in making payments, so that it can be indirectly introduced into the utility and production functions. These two components of this approach are presented in this chapter.
2. Money is not directly or indirectly in the utility and production functions but is required for certain types of transactions, so that a cash-in-advance analysis becomes appropriate. This cash-in-advance approach can be found in Chapter 23.
3. Money is not directly or indirectly in the utility and production functions and is not used as a medium of payments in a cash-in-advance manner. However, money is an asset that

2 Intuitively, commodities are goods directly used in consumption or production. Financial assets are paper or book-keeping claims to commodities and are used for their liquidity services or to transfer purchasing power from the present to the future.

can be used for transferring purchasing power across periods. This approach is used in the overlapping generations models of money, presented in Chapters 21 and 22.

Of these three approaches, the most common is the first; money can be treated as a component of the utility and production functions.

Money in the utility function and the production function

Our preferred approach to money in this chapter puts it in the individual's utility function and the firm's production function because it is the medium of payments in a monetary economy in which commodities (bonds) do not trade against other commodities or bonds but do so only against money. This approach is known as the money in the utility function (MIUF) and the money in the production function (MIPF) approach. Many economists object to this approach on the grounds that real balances do not "directly yield satisfaction or increase production." An indirect route to this approach is a "transactions" approach that initially keeps money out of the utility and production functions; however, the use of money allows the consumer to reduce transactions time for payments and therefore to increase leisure by using money, and its use also allows the firm to save on its labor resources. These arguments lead to the *indirect* utility and production functions, which are briefly presented in Sections 3.3.2 and 3.6.2 of this chapter.

However, many economists prefer to completely eschew the above lines of analyses, with some of them opting for money in an overlapping generations framework. This approach is presented in Chapters 21 to 23.

Money as a durable good

Financial assets are durable goods in an economic sense. The concept of the economic durability of money can be quite confusing and needs clarification.

The demand for money is taken to be a demand for the *average money balances held by the individual in a period* and is often designated as *the demand for nominal balances to hold*. This demand differs from the amounts that the individual would hold at various points in time during the period but is a weighted average of the latter amounts, with the weights being the duration a particular amount is held.[3]

However, an individual may or may not hold a durable good for its transactions services. He may instead use it as a means of transferring his wealth or real purchasing power from

3 To consider an example, assume that an individual holds \$100 at the beginning of a week and spends it at a continuous even rate over the week. His average money balances – designated as his demand for money – held over the month are \$50 (= 100/2) which clearly differ from his money balance of \$100 at the beginning of the week and his money balance of zero dollars at the end of the week. For comparable period analysis, assume that he spent \$100/7 (=\$14.29) per day of the week. He would then hold \$85.71 (=100 – 14.29) for 6 days, \$71.42 (= 85.71 – 14.29) for 5 days, and so on. The weighted average (i.e. weighted by the number of days held) of these amounts would be \$42.86, so that there is a slight difference between the continuous and the discrete cases for the average calculation. We will proceed with the continuous even expenditure assumption, which implied the weighted average balance to be \$50. Under this assumption, the individual would be taken to have had an average demand for \$50 of money balances, a durable good, and to have used its services in financing his purchases during the week. See also Chapter 4 on this point.

one week to the next.[4] Such a usage would be one of a *store of value*.[5] For convenience, monetary theory has generally treated the demand for money as a medium of payments under the category of the transactions demand for money and the demand for money as a store of value (relative to other assets) as the speculative or portfolio demand for money. But any particular unit of money balances can be used for either function, and the division into the transactions and speculative balances must be taken to be an analytical division and not necessarily applicable to the real world. This chapter confines itself to general propositions on the total demand for money.

3.2 Stylized facts of a monetary economy

As pointed out at many points earlier in this book, the essential role of money is that of medium of payments. To perform this role, it needs to be a store of value, at least over short intervals from receipt of money to its payment to others. The macroeconomic definition of bonds is non-monetary financial assets. Such assets also function as stores of value, often better than money since they usually provide higher returns than money. What are the main stylized facts related to money in a modern economy that a theory that purports to have money in it must satisfy? Our simple and short list of these stylized facts on money is as follows:

1 Commodities, labor and bonds do not exchange against each other but only against money.
2 The income from the supply of labor or accruing in other ways is received in money, while the purchases of commodities and bonds have to be paid for in money. Since these two actions do not occur at the same instant, money is held in every period (which is long enough to include both the receipt of income and expenditures from it). By its very nature as a store of value, it can also be held from one period to the next. Therefore, in a monetary economy, there is a positive demand for money in every period.[6]
3 The demand for money is positive, irrespective of whether the return on it is higher or lower than the return on bonds. In fact, the return on money is usually less than on bonds, but money demand is nevertheless positive. The positive demand for money as the medium of payments coexists with a positive demand for both risky and riskless bonds.[7]

4 A pure store of value without any transactions usage would occur if the individual held $50 consistently – and never spent any of it – from the beginning of the week to the end of the week. He would then have bequeathed this amount to the beginning of the following week, much in the manner of a durable consumer good such as a refrigerator, which outlasts the current week of usage and is still available to the individual at the beginning of the following week. Thus, for the pure store-of-value function, the individual could store the unplugged refrigerator or the $50 of money balances through the week without any intention of using their services for refrigeration or financing payments, respectively. In practice, both the refrigerator and the money balances will see some usage – the latter for financing transactions – during the week and still act as stores of value. Chapter 4 presents the analysis of the transactions usage combined with the store of value role of money.
5 Friedman called the temporary store of value for which money is used an abode of purchasing power.
6 This implies that in any model involving more than one period money needs to have a strictly positive demand in all periods, including both the first period and the last period of the analysis.
7 If a model implies that both the demand for money and for bonds are not simultaneously positive, it is the demand for bonds that has to be zero.

4 The demand for money is positive, irrespective of whether the return on it is higher or lower than the return on commodities through storage or production of other commodities. It is also positive even if a positive rate of inflation is expected.
5 For individuals, the demand for money is a function of total expenditures or of consumption expenditures, not of saving. In particular, it can be either greater than or less than saving, but virtually never equals the saving in a given period.
6 The velocity of circulation of money is positive over periods that include both the receipt of income and the purchases of commodities, but is nevertheless not constant.

Students should check the validity of points 1 to 5 with their own behavior. For many of them, expenditures are equal to or below their incomes. Therefore, they dissave, with the dissaving financed by their issue of bonds (IOUs to parents, loans from the universities and the government, etc.). In spite of zero or negative saving, they hold positive amounts of money (in fact, both currency and demand deposits). Chapters 14, 21 and 23 provide longer lists of the stylized facts on money in the economy.

Macroeconomics and monetary economics provide several models with money listed as a variable in the model. It is often designated, by assumption, as an asset that is riskless and has a zero return. However, neither of these is among the essential characteristics of money, so that including an asset with these characteristics and calling it "money" does not mean that there really is money in the model. Money would be a misnomer for such an asset unless it meets the preceding list of stylized facts. In short, our intention is to use these facts to discriminate (reject or accept) among models that include an asset that they call "money."

The next section presents a commonly used macroeconomic model that claims to include money. However, its implications for the demand for money run foul of the stylized facts, so we argue that there is really no money in it, which leads to its rejection as a valid model for a monetary economy. Chapters 21 to 23 present some OLG models with "money." The benchmark model of this approach, specified in Chapters 21 and 22, also fails to satisfy the stylized facts on money.

3.3 Optimization without money in the utility function

As discussed above, the essential role of money is that of medium of payments. To perform this role, money also has to be a store of value, at least for short durations, or, as Milton Friedman put it, "a temporary store of value." What is the appropriate model of consumer behavior for capturing these roles? This section motivates discussion on this issue by the use of a standard two-period model, without uncertainty, of consumer behavior for an economy with commodities, money and bonds.

Assume that, in period 1, the individual has the frequently used two period utility function of the form:

$$U(c_1, n_1, c_2, n_2) \tag{1}$$

where c_i and n_i are respectively the consumption of commodities and the supply of labor in the ith period. $U(.)$ is assumed to be an ordinal neoclassical utility function with continuous first- and second-order partial derivatives. Note that money does not appear in this

utility function. For simplification, assume that the utility function has the common time-separable form, so that:

$$U(c_1, c_2) = u(c_1, n_1) + \frac{1}{1+\rho} u(c_2, n_2) \tag{2}$$

where ρ is the rate of time preference and u is the period utility function. $\frac{\partial u_i}{\partial c_i} > 0$ and $\frac{\partial u_i}{\partial n_i} < 0$ for all i.

In each period, the individual uses his nominal income $P_1 y_1$ plus the "inherited" (i.e. from the preceding period) amounts of money and bonds to buy commodities, money and bonds. Commodities are wholly consumed during the period whereas money and bonds are carried to the next period. Money does not pay interest, but bonds pay a nominal interest rate R per period. The individual's budget constraint for period 1 is:

$$P_1 c_1 + M_l + B_1 = P_1 w_1 n_1 + M_0 + (1 + R_0) B_0 \tag{3}$$[8]

where M is nominal money balances, B is the nominal value of bonds and w is the exogenously given real wage rate. At the beginning of period 2, the individual has the carryover money balances of M_1 (which do not pay interest), carryover bonds of B_1 paying interest at the rate R, and receives income y_2. In a two-period model without a bequest motive, the individual will not buy money and bonds in period 2 since they are of no use to him after the end of that period, so purchases of the two assets do not appear in the budget constraint for period 2. With $M_2 = B_2 = 0$, the second-period budget constraint is:

$$P_2 c_2 = P_2 w_2 n_2 + M_1 + (1 + R_1) B_1 \tag{4}$$[9]

Since the individual is not able to issue money, we also have:

$$M_1 \geq 0 \tag{5}$$

Solving (4) for B_1 and substituting in (3), the consolidated budget constraint for the two periods is:

$$P_1 c_1 + \frac{P_2 c_2}{1+R_1} + M_1 - \frac{M_1}{1+R_1} = p_1 w_1 n_1 + \frac{P_2 w_2 n_2}{1+R_1} + M_0 + (1+R_0) B_0 \tag{6}$$[10]

which yields:

$$P_1 c_1 + \frac{P_2}{1+R_1} c_2 + \frac{R_1}{1+R_1} M_1 = P_1 w_1 n_1 + \frac{P_2}{1+R_1} w_2 n_2 + M_0 + (1+R_0) B_0 \tag{7}$$

8 We have specified this constraint as an equality for the rational individual since he would either consume all his endowments or convert any saving into money m for possible use in the future.

9 Since the individual derives no utility from unspent money balances or unconsumed commodities left over at the end of $t+1$, utility maximization implies that the constraint is an equality.

10 We have specified this constraint as an equality for the rational individual since he would consume all his endowments over the two periods.

Dividing through by P_1 and substituting $1/(1+r_1)$ for $P_2/(P_1(1+R_1))$, we have:

$$c_1 + \frac{c_2}{1+r_1} + \frac{R_1}{1+R_1}\frac{M_1}{P_1} = w_1 n_1 + \frac{w_2 n_2}{1+r_1} + \frac{M_0}{P_1} + \frac{(1+R_0)B_0}{P_1} \tag{8}$$

where $1/(1+r_t) = P_{t+1}/[P(1+R_t]$, r_t is the real rate of interest in period t and P_{t+1}/P_t is the inflation rate.

Replacing M/P by m (real money balances) and B/P by b (real value of bonds), and simplifying, (8) can be restated as:

$$c_1 + \frac{c_2}{1+r_1} + \frac{R_1}{1+R_1} m_1 = w_1 n_1 + \frac{w_2 n_2}{1+r_1} + \frac{P_0}{P_1} m_0 + (1+r_0) b_0 \tag{9}$$[11]

The real value of money and bonds at the beginning of period t are the "endowments" of period t. Replacing $(P_0/P_1)m_0 + (1+r_0)b_0$ by a_1, we have:

$$c_1 + \frac{c_2}{1+r_1} + \frac{R_1}{1+R_1} m_1 = w_1 n_1 + \frac{w_2 n_2}{1+r_1} + a_1 \tag{10}$$

The individual maximizes (2) subject to (10), so that the Lagrangean function, with the Lagrangean multiplier λ, is:

$$L = u(c_1, n_1) + \frac{1}{1+\rho}\, u(c_2, n_2) + \lambda\left(w_1 n_1 + \frac{w_2 n_2}{1+r_1} + a_1 - c_1 - \frac{c_2}{1+r_1} - \frac{R_1}{1+R_1} m_1\right) \tag{11}$$[12]

subject to $m_1 \geq 0$.

The first-order conditions are:

$$\frac{\partial L}{\partial c_1} = \frac{\partial u(c_1, n_1)}{\partial c_1} - \lambda = 0 \tag{12}$$

$$\frac{\partial L}{\partial n_1} = \frac{\partial u(c_1, n_1)}{\partial n_1} + \lambda w_1 = 0 \tag{13}$$

$$\frac{\partial L}{\partial c_2} = \frac{1}{1+\rho}\frac{\partial u(c_2, n_2)}{\partial c_2} - \frac{\lambda}{1+r_1} = 0 \tag{14}$$

$$\frac{\partial L}{\partial n_2} = \frac{1}{1+\rho}\frac{\partial u(c_2, n_2)}{\partial n_2} + \frac{\lambda w_2}{1+r_1} = 0 \tag{15}$$

$$\frac{\partial L}{\partial m_1} = -\lambda\frac{R_1}{1+R_1}, \quad m_1 \geq 0, \quad m_1\frac{\partial L}{\partial m_1} = 0 \tag{16}$$

The first four of these conditions are identical to those in a model without money (though there may be bonds) in the economy, so that their solution provides the optimal time path of

11 Note that
$$\frac{(1+R_0)P_0}{P_1} b_0 = (1+r_0)b_0.$$

12 Note that
$$\frac{(1+R_0)P_0}{P_1} b_0 = (1+r_0)b_0.$$

consumption and labor supply, which is independent of the existence or absence of money in the model. Money is not only neutral (i.e. the invariance of the real values of the real variables of the model with respect to changes in the nominal quantity of money) in this model, it is "*strongly neutral*"[13] in the sense that the optimal real values of the real variables (consumption and labor supply) of the model are not only invariant with respect to changes in the quantity of money, but are also invariant with respect to its existence versus absence in the economy. That is, the time path of consumption in this model is the same as in an economy with bonds but no medium of payments. While there are considerable disputes among economists on the empirical validity of the neutrality of money, no one subscribes to the strong super-neutrality proposition since real-world monetary economics do change, often drastically, the values of the real variables relative to a barter economy or one with bonds but no medium of payments.[14] This would be even more apparent if the labor supply in the two periods were also inserted in the utility function.[15] In this extension, even the labor supply and leisure in each period would also be invariant to the existence of money, so that their demand will be the same in a monetary as in a barter economy. This is also patently invalid for real-world economies.

Looking at the last of the first order conditions above, if $R_1 > 0$, $-R_1/(1+R_1) < 0$. Since, from (11), $\lambda > 0$, $-\lambda/R_1/(1+R_1) < 0$, so, by the slackness condition, $m_1 = 0$. In this scenario, bonds have a higher return than money, which does not pay a positive return, so that bonds are held but money is not held. But if $R_1 < 0$, $-\lambda/R_1/(1+R_1) > 0$ so the individual will hold as much money as possible. If he cannot borrow, he will hold money in period 1 *equal to his saving* in that period. If he can borrow, his money holdings will equal his saving plus his borrowing. In this case, money balances pay a higher return than bonds, so that while money is held, bonds are not held. To conclude:

$$\text{If } R_1 > 0, \quad m_1 = 0.$$

$$\text{If } R_1 \leq 0, \quad m_1 \geq 0.^{16}$$

Since R_1 is the market rate on bonds, the individual will have no incentive to lend if $R_1 \leq 0$, so any saving will be held in money. Hence, in the usual scenario with $R_1 > 0$, the individual will not hold money in period 1 as well as in period 2, but will hold bonds, equal to his saving, in period 1, though not in period 2. On the demand for money, students can consider their own behavior pattern. They often have incomes below their expenditures, with the dissaving financed by the issue of bonds (IOUs to parents, loans from the universities and the government, etc.). Nevertheless, they do hold money. Therefore, for them, $m_1 > 0$ while $b_1 < 0$. Such behavior refutes the above implications of the model of this section.

For the preceding model, our analysis derived several implications that are clearly (empirically) invalid for monetary economies, as can be seen by comparing this model's implications with the stylized facts of monetary economies listed in the preceding section. One of these is the strong super-neutrality result. A second one is that the demand for money is not positive in every period: in the model, either money or bonds is held, but not both simultaneously, and in the realistic case of $R_1 > 0$ money is not held. This is also clearly invalid for monetary economies in which each individual does hold money in every period,

13 The definitions of neutrality and super-neutrality are given later in this chapter.

14 More detailed discussion of this point is given in Chapter 24.

15 This is left as an exercise for students.

16 The equality holds if saving is zero.

whether or not he also holds bonds. The explanation for this result is that, in the above model, both money and bonds perform exactly the same role; they are both stores of value, but neither is a medium of exchange. However, in monetary economies, money is the medium of payments, bonds are not, so that money is different in nature from bonds. The preceding model does not capture this distinction and, in fact, has no medium of payments. The "money" it has is purely a store of value, just as bonds are, and could have been equally well designated as zero-interest bonds. The medium-of-payments role of money arises from the existence of and trade among numerous commodities within *each* period. This does not occur in the above model.

A third invalid implication of the above model is that for $R_1 > 0$, $m_1 = 0$. However, in a monetary economy, the individual receives the income from his labor services in the form of money. Since there is realistically always an interval between the receipt of this money and its payment to others for commodities or the investment of saving in bonds, money will be held, so that its demand in both periods 1 and 2 cannot be zero, even if the return on bonds exceeds that on money. A fourth invalid implication of the preceding model is that the individual's money holdings in the last period, m_2, equal zero. In reality, in monetary economies, individuals in the last period of their lives do hold money since they receive it in exchange for labor supply and somewhat later buy commodities for consumption, i.e. they hold it for its medium-of-payments role even if they have no bequest motive. In fact, the challenge of monetary theory can be said to show that money is held even if bonds dominate it in return; further, it is held in each period, including the last one, when there is no future use for a store of value. The above model fails this challenge.

Of interest in a university class is another invalid implication of the model that occurs when there is dissaving in period 1. The model in this case implies that money demand will be zero. Most university students' behavior and circumstances are such that they spend more on commodities than their income, thereby dissaving and accumulating debt (i.e. negative bond holdings), while they continue to have positive desired money holdings. At the other end of lifetime, seniors in their last year of life continue to buy commodities and hold positive desired money holdings in order to make their purchases.

Given the empirical invalidity of several core implications of the preceding model for money as a medium of payments, the following sections replace it by two others that do imply a positive demand for money as a medium of payments. To show that such a demand can arise within each period rather than because of the use of money as a way of transferring purchasing power (store of value) across periods, we shall use a one-period model with heterogeneous commodities and labor supply, though the analysis can be readily adapted to a two-period or multi-period format, as is done in Chapter 24 in the context of overlapping generations models. The models used in this chapter are the money in the utility function (MIUF) model and the money indirectly in the utility function (MIIUF) model. Another approach is that of cash-in-advance models. This approach is omitted from this chapter but is presented in Chapter 24 in the context of overlapping generations models.

3.4 Medium of payments role of money: money in the utility function (MIUF)

As discussed in the introduction to this chapter, money balances can be inserted as a variable in the utility function or in the indirect utility function. Since doing so is a matter of some dispute in the literature, this section presents the justifications for doing so at somewhat greater length than is normally done in monetary economics textbooks.

As discussed at the end of the previous section, the study of the role of money as a medium of payments does not require multi-period analysis which, by focusing on the role of money as a store of value similar to bonds, detracts from its role as a medium of payments. Consequently, we will use one-period analysis. Further, our introduction of money in the utility function is at its core the introduction of money's liquidity services as a medium of payments. These services occur in a monetary economy, but not in a barter one, because of the *environment* of a monetary economy, which is that, in such an economy, commodities and bonds trade against money and not against other commodities and bonds. This environment creates a preference for a larger rather than a smaller amount of real balances, until satiation in them is reached, so that the medium-of-payments services rendered by money holdings become one of the arguments of the utility function of individuals buying and selling commodities in a monetary economy. These services are a real good and can be proxied by the amount of real balances held. Consequently, the assets that do not yield liquidity services in facilitating transactions during the current period will be excluded from the utility function. Stocks and long-term bonds are among such assets. However, short-term bonds and savings and term deposits do possess some liquidity in modern economies and are often taken to be near-monies. This leads to questions about the definition of money. This was partly addressed in Chapter 1 and will be more fully dealt with in Chapter 7 on monetary aggregation. For the time being, "money" will be understood to include financial assets possessing liquidity and, with further simplification, will be taken to be M1.

3.4.1 Money in the utility function (MIUF)

This subsection presents the axiomatic basis for including money in the utility function.

Individuals differ in their tastes or preferences over goods and in their income or wealth. Microeconomic theory defines the "rational" individual as one whose preferences are consistent and transitive.[17] The definitions of these terms are specified by the following axioms of utility theory:

Axiom (i): Consistent preferences
If the individual prefers a bundle of goods A to another bundle B, then he will always choose A over B.
Axiom (ii): Transitive preferences
If the individual prefers A to B and B to a third bundle of goods C, then he prefers A to C.

To these two axioms in the theory of the demand for commodities, monetary theory usually adds the following one:

Axiom (iii): Real balances as a good
In the case of financial goods that are not "used directly in consumption or production" but are held for exchange for other goods in the present or the future, the individual is concerned

17 An additional axiom is sometimes added for analytical convenience. This is that the individual never reaches satiation for any good. That is, he continues to prefer more of each good to less of it. In view of the definition of goods above, this axiom implies that a good never ceases to be a good for the individual, no matter how much or how little he possesses of it.

with the former's exchange value into commodities – that is, their real purchasing power over commodities and not with their nominal quantity.[18]

The axioms of consistency and transitivity ensure that the individual's preferences among goods can be ordered monotonically and represented by a utility or preference function. Axiom (iii) ensures that financial assets, when considered as goods in such a utility function, should be measured in terms of their *purchasing power* and not their nominal quantity. The inclusion of money – and other financial assets – directly into the utility function can be justified on the grounds that the utility function expresses preferences and that, since more of financial assets is demanded rather than less, they should be included in the utility function just like other goods.

Given these axioms, let the individual's period utility function be specified as:

$$U(.) = U(x_1, \ldots, x_K, n, m^{\mathrm{h}}) \tag{17}$$

where:

x_k = quantity of the kth commodity, $k = 1, \ldots, K$
n = labor supplied, in hours
m^{h} = average amount of real balances held by the individual or household for their liquidity services.

Note that (17) has K+2 goods, consisting of K commodities, labor and real balances.

Axioms (1) to (3) only specify $U(.)$, an ordinal utility function.[19] $U_k = \partial U/\partial x_k > 0$ for all k, $U_n = \partial U/\partial n < 0$, $U_m = \partial U/\partial m^{\mathrm{h}} > 0$. All second-order partial derivatives of $U(.)$ are assumed to be negative. That is, each of the commodities and real balances yield positive marginal utility and hours worked have negative marginal utility.

The complete MIUF model is presented after the next sub-section.

3.4.2 Money in the indirect utility function (MIIUF)

It is sometimes asserted that money does not directly yield consumption services to the individual, but that its use saves on the time spent in making payments. This first part of this assertion implies that the first two axioms of preferences in the preceding subsection are not applied to real balances but only to commodities and leisure.

A model that leaves real balances out of the direct utility function but embodies their usage for facilitating purchases and sales of commodities is briefly specified in this subsection. For this model, assume that only consumer goods and leisure directly yield utility. Hence, the one-period utility function $U(.)$ is:

$$U(.) = U(c, L) \tag{18}$$

18 Thus, 100 bank notes each with a face value of \$1 have a nominal quantity or value of \$100. Assume that the individual wishes to hold a certain amount of real purchasing power in money balances and this demand of his equals \$100 at a certain set of prices. If prices of commodities were to double, the individual would no longer demand \$100 but \$200 of money balances in order to keep his demand constant in terms of real purchasing power.

19 A utility function that gives a consistent and transitive ranking of preferences, without any other characteristics of measurability, is said to be ordinal or unique up to an increasing monotonic transformation. That is, if $U(x1, \ldots, xs)$ is the individual's utility function, then $F[U(x1, \ldots, xs)]$, where $\partial F/\partial U > 0$, is also an admissible utility function with identical demand functions for xi, $i = 1, \ldots, s$.

where:

c = consumption
L = leisure.

Assume that $U_c, U_L S > 0, U_{cc}, U_{LL} > 0$. Consumption requires purchases of consumer goods, which necessitate time for shopping. This shopping time can be divided into two components, one being the selection of the commodity to be purchased and the other that of making the payment acceptable to the seller. The former is often enjoyable to most people and can be treated as an aspect of the commodity bought, or as a use of leisure, or ignored as a simplification device for our further analysis. The second component is an aspect of the payments system. If the buyer does not have enough of the medium of payments to pay for the purchase, he has to devote time to getting it, say, from a bank, or to find a seller who will be willing to accept the payment in the commodity or labor services that the seller can provide, where the latter is the time taken by bartering. Both of these clearly take time. In a monetary economy, over all his purchases, the buyer needs a certain amount of money to buy all the goods and services that he wishes to purchase. He can hold enough or only some proportion of this amount. If he holds less than 100 percent of the amount needed, he will have to devote part of his time to effect the remaining payment by devoting some time to the payments process. The amount of time needed for this purpose will be positively related to the shortfall in his money holdings. The time used for this purpose is a nuisance, would have negative marginal utility and can be labeled as "payments time" – that is, the time needed to effect the payments for the commodities bought. It is also often labeled as "shopping time" or "transactions time."

Leisure equals the time remaining in the day after deducting the time spent on a job and the payments time. Hence,

$$L = h_0 - n - n^{\mathrm{T}} \tag{19}$$

where:

h_0 = maximum available time for leisure, work and transactions
n = time spent working
n^{T} = payments time, i.e. time spent in making payments in a form acceptable to the seller.[20]

The payments and financial environment are assumed to be such that the "*payments/transactions time function*" is:

$$n^{\mathrm{T}} = n^{\mathrm{T}}(m^{\mathrm{h}}, c) \tag{20}$$

where $\partial n^{\mathrm{T}}/\partial c > 0$ and $\partial n^{\mathrm{T}}/\partial m^{\mathrm{h}} \leq 0$. From (19) and (20), $\partial U/\partial n^{\mathrm{T}} = (\partial U/\partial L)(\partial L/\partial n^{\mathrm{T}}) < 0$. That is, an increase in payments time decreases leisure and therefore decreases utility. But, since an increase in the amount held and utilized of real balances decreases payments time, $\partial U/\partial m^{\mathrm{h}} = (\partial U/\partial n^{\mathrm{T}})(\partial n^{\mathrm{T}}/\partial m^{\mathrm{h}}) > 0$.

20 Since this transactions time reduces leisure and leisure has positive marginal utility, transactions time in this model has negative marginal utility. This is quite reasonable since we are considering the transactions time made necessary by not having adequate money to pay for one's purchases, rather than transactions with enough money in hand to pay for the desired level of purchases. The latter may be enjoyable, while the former is likely to be the chore of finding sellers who will transfer their goods in some form of barter.

Equation (20) specifies the time it takes to pay for an amount c of commodities while utilizing an average amount m^{h} of real balances. In a monetary economy in which the shops would only sell against money, the time required to pay for any positive level of commodities would become infinitely large as the individual tries to do without money. That is, as $m^{\mathrm{h}} \to 0$, $n^{\mathrm{T}} \to \infty$. For positive levels of real balances, $\partial n^{\mathrm{T}}/\partial m^{\mathrm{h}} \leq 0$. The reason for this is that, in a monetary economy, money is the most widely accepted medium of payments, so that trying to pay in any other way may mean searching for special suppliers, which would increase the payments time.[21] However, beyond some limit, say for $m^{\mathrm{h}} \geq \alpha c$, where α is the inverse of the velocity of circulation of money applicable to the individual, there is unlikely to be any further decrease in payments time from additional real balances, so that, beyond this limit, $\partial n^{\mathrm{T}}/\partial m^{\mathrm{h}} = 0$.

A proportional form of the payments time function is:

$$n^{\mathrm{T}}/c = \phi(m^{\mathrm{h}}/c) \tag{21}$$

where $-\infty < \phi' \leq 0$, with ϕ' as the first-order derivative of ϕ with respect to m^{h}/c. Satiation in real balances occurs as $\phi' \to 0$. (21) implies that $\partial\phi/\partial m^{\mathrm{h}} \leq 0$. Incorporating this payments time function into the utility function above, we have:

$$U(.) = U(c, h_0 - n - c\phi(m/c)) \tag{22}$$

(22) can be rewritten as the indirect utility function:

$$V(.) = V(c, n, m^{\mathrm{h}}) \tag{23}$$

where $\dfrac{\partial V}{\partial m^{\mathrm{h}}} = \dfrac{\partial U}{\partial L}\left[-c\dfrac{\partial \phi}{\partial m^{\mathrm{h}}}\right]$. Since $\dfrac{\partial U}{\partial L} > 0$ but $\dfrac{\partial \phi}{\partial m^{\mathrm{h}}} \leq 0$, $\dfrac{\partial V}{\partial m^{\mathrm{h}}} \geq 0$.

The generic form and properties of the indirect utility function (23), which has real balances as a variable, are similar (though not identical[22]) to those of the direct one used earlier in this chapter. Therefore, economists who prefer its payments time justification for putting money in the utility function substitute this justification for the one given earlier for the direct MIUF, which was simply that money is in the utility function because the individual prefers more of it to less, *ceteris paribus,* in the environment of a monetary economy. Both justifications

21 We illustrate this by an "island parable" in which the sellers of a commodity are located on different islands, with each seller having only one unit of the commodity to sell. Assume that all sellers are willing to accept money in payment for the purchases of the commodity while only some sellers are willing to sell their commodity in exchange for some other means of payments (transfer of bonds, IOUs, or other commodities). Further, assume that each buyer needs to buy c units of the commodity but does not know the island locations of the sellers *who will accept only money in payment* and must search on a random basis among all islands. This buyer needs to visit c islands (since only one unit of the commodity can be bought on each island) with sellers who will accept the means of payment carried by the buyer. To buy c units of the commodity, a buyer who carries enough money balances will have to visit c islands, while other buyers with less than c units of money balances will have to visit more islands and spend more time in the search process.

22 For example, the direct utility function has $\partial U/\partial m \to \infty$ as $m \to 0$, while this need not occur in the indirect utility function. For an example of such a condition, let the transactions time without money reach a finite constant ck, so that we have $\partial\phi/\partial m \to ck$ as $m \to 0$. Then, we do not have $\partial V/\partial m \to \infty$ as $m \to 0$. The indirect utility function would have satiation in money holdings at $m \geq c$. However, such conditions can be derived from the latter function and imposed on the former.

are acceptable. However, given the similarity of the direct and the indirect utility functions, and the relative simplicity of using the former, we revert for convenience to the direct utility function.

3.4.3 Empirical evidence on money in the utility function

Money directly or indirectly in the utility function seems to provide the most realistic results on the demand for money and the relationship between money and output, as compared with models which do not do so (see the stylized facts on money in this chapter and Chapters 14 and 21).

More directly, on estimation of the utility function itself, a specific form of the utility function is the constant elasticity of substitution function (see Chapter 7):

$$U(c_t, m_t) = [\alpha c_t^{1-v} + (1-\alpha) m_t^{1-v}]^{1/(1-v)} \quad 0 < 1, v > 0, v \neq 1 \tag{17'}$$

For $v = 1$,

$$u(c_t, m_t) = c_t^{\alpha} m_t^{1-\alpha} \tag{17''}$$

where (17″) is the Cobb–Douglas form with the elasticity of substitution between c and m being unity.[23] Holman (1998) reports estimates, based on US data from 1889 to 1991, of v around 1.0 and of α around 0.95, while shorter periods fail to reject $v = 1$. Other estimates of these parameters can be deduced from studies estimating the demand for money, which are covered in Chapter 9. In any case, empirical studies do not reject the notion that money can be treated as a component of the utility function.

3.5 Different concepts of prices

Prices, like temperature, distance, etc., have to be measured in terms of a scale. Such a scale for measuring prices is called a *unit of account*. The goods that serve as a medium of payments in a certain society may or may not be actually used as a unit of account in that society or may only do so for certain purposes.[24]

The prices of individual goods measured in terms of a unit of account are referred to as *accounting prices*.[25] If the unit of account is money, then the prices are implicitly in terms of

23 Walsh (2003, Ch. 2) provides more extensive forms of utility functions with money as an argument.

24 In economies with hyperinflation in terms of the domestic currency, one or more foreign currencies, or gold, are often used as the unit of account.

25 From a rigorous viewpoint, money prices are accounting prices. However, the convention has grown up in economics that only prices measured in terms of a unit that is not one of the goods in the economic system itself are called accounting prices and only the nongood unit is called a unit of account. Such a unit of account is, for example, the guinea in England. The guinea has no physical counterpart in the real world. Traditionally, it had the value of £1.05 – which used to correspond to the old 21 shillings, while the pound was worth 20 shillings. Assume, however, that its value was halved by a decree or fiat to £0.0525. Since the guinea has no real existence and is not demanded or supplied in the economy, the halving of the value of the guinea in terms of the monetary units of pounds would not affect behavior in the economy. Each money price – that is, the price of a good in terms of money – would remain the same. However, each accounting price calculated in terms of the guinea would double. Such a change in all accounting prices – or, as expressed alternatively, a change in the value of the nongood unit of account – does not affect the quantities demanded or supplied or the monetary prices of goods.

money but are sometimes more explicitly referred to as "*money prices*," "*monetary prices*," "*absolute prices*" or "*prices in terms of money*."

If prices are measured in terms of money, then the price of a nominal unit of money itself must be unity, since a dollar note has a price of one in terms of itself. Hence, the *price of nominal balances* is a constant at unity and cannot change. However, the *price* (of a unit) *of real balances* is the price level itself.

The term "*price level*" or "*general price level*" is the weighted average of the prices of a representative bundle of the commodities in the economy. The price level is, in practice, measured by an index whose mode of calculation is specified later by equation (34).

3.6 User cost of money

For one-period analysis, the cost of using the services of a durable good or asset is its rental or user cost during the period. This cost is the sum of the interest cost and depreciation less the increase in the capital value of the good during the period. This is also the relevant concept for using the services of money in facilitating exchanges among commodities.

The user cost of real balances is specified as the *interest foregone from holding real balances* relative to holding a totally illiquid asset.[26] That is, the user cost ρ_m of *real* balances[27] is:

$$\rho_m = (R - R_m)P \tag{24}$$

and the user cost ρ'_{M} per unit of *nominal* balances is:

$$\rho'_{\mathrm{M}} = (R - R_m) \tag{24'}$$

where:

ρ_m = nominal user cost per unit of real balances
R = nominal/market interest rate on the illiquid asset
R_m = nominal interest rate paid on nominal balances
P = price level
m = real balances.

In (24) and (24′), $(R - R_m)$ is the interest foregone from holding a dollar of nominal balances. On an amount m of real balances, Pm is the nominal value of the m real balances and $(R - R_m)Pm$ would be the total rental cost of these balances.

This is hardly surprising since nongood units of account are bookkeeping devices and do not affect economic behavior.

26 This is the usual way of specifying the user cost of money. However, in the real world, the user cost of money often does have additional components such as the time taken to obtain cash balances, etc. These are more explicitly considered in Chapter 4 on the transactions demand for money.

27 The user cost over a period of a unit of the ith durable commodity is $(r + d - \pi_i)p_i$, where R is the market rate of interest, d is the rate of depreciation of the commodity, π_i is the rate of increase in the price of the ith commodity and p_i is its price. Applying this formula to the case of money, d is zero; in perfect financial markets, the rate of interest R incorporates the rate of inflation π for all commodities and the price of nominal balances is unity.

3.7 The individual's demand for and supply of money and other goods

3.7.1 Derivation of the demand and supply functions

To derive the individual's demand and supply functions for all goods, maximize:

$$U(x_1, \ldots, x_K, n, m^{\mathrm{h}}) \tag{25}$$

subject to:

$$\Sigma_k p_k x_k + (R - R_m) P m^{\mathrm{h}} = A_0 + Wn \quad k = 1, \ldots, K \tag{26}$$

where:

p_k = price of kth commodity
P = price level
W = nominal wage rate
A_0 = nominal value of initial endowments of commodities and financial assets.

Maximizing (25) subject to (26) gives the first-order maximizing conditions as:

$$U_k - \lambda p_k = 0 \quad k = 1, \ldots, K \tag{27}$$

$$U_n + \lambda W = 0 \tag{28}$$

$$U_m - \lambda (R - R_m) P = 0 \tag{29}$$

$$\Sigma_k p_k x_k + (R - R_m) P m^{\mathrm{h}} = A_0 + Wn \tag{30}$$

where λ is the Lagrangean multiplier. Equations (27) to (30) constitute a system of $K + 3$ equations in the $K + 3$ endogenous variables $x_1, \ldots, x_K, n, m^{\mathrm{h}}$ and λ. The exogenous variables are: $p_1, \ldots, p_K$, W, R, R_m and P.

Assuming that a unique solution exists for the set of equations (27) to (30) and that the sufficiency conditions for a maximum are satisfied, the solution for the K+3 endogenous variables will have the general form:

$$x_k^{\mathrm{dh}} = x_k^{\mathrm{dh}}(p_1, \ldots, p_K, W, (R - R_m)P, A_0) \quad k = 1, \ldots, K \tag{31}$$

$$n^{\mathrm{s}} = n^{\mathrm{s}}(p_1, \ldots, p_K, W, (R - R_m)P, A_0) \tag{32}$$

$$m^{\mathrm{dh}} = m^{\mathrm{dh}}(p_1, \ldots, p_K, W, (R - R_m)P, A_0) \tag{33}$$

where the superscripts d and s stand for the demand and supply functions respectively and the superscript h stands for households.

3.7.2 Price level

The price level P is related to $p_1, \ldots, p_K$ by the index number formula:

$$P_t = [\Sigma_\kappa \, p_{kt} \, x_{k0}] / [\Sigma_\kappa \, p_{k0} \, x_{k0}] \tag{34}$$

where the subscript t refers to the period t and the subscript 0 refers to the base period for the construction of the price index. $x_{k0}, k = 1, \ldots, K$, is the weight attached to the kth commodity

used in constructing the price index and is usually specified as the amount of the commodity purchased in the base period.

A common example of a price index constructed according to (34) is the Consumer Price Index (CPI). For such an index, x_{k0} is the amount of commodities bought for consumption in the economy during the base year.

Another popular price index is the GDP (Gross Domestic Product) deflator. The latter takes the composite bundle of commodities to be used for (34) as the commodities included in GDP, with their weights specified by their weight in GDP. The GDP deflator includes both capital and consumer goods, while the CPI excludes capital goods. Since our concern will usually be with the total output of the economy, for our purposes the GDP deflator will be the more appropriate proxy for the theoretical concept of the price level.[28]

Our main concern with the price index will be with its homogeneity properties. Equation (34) has the property that the price level is homogeneous of degree one in all prices, so that a doubling of the latter will double the former also. That is,

$$\alpha P_t = (\alpha p_{1t}, \ldots, \alpha p_{Kt}) \quad \text{for } \alpha > 0 \tag{35}$$

3.7.3 Homogeneity of degree zero of the demand and supply functions

We can now determine the effects on the individual's demand and supply functions of increasing the nominal variables $p_1, \ldots, p_K$, W and A_0 by an identical proportion, such that these values are replaced respectively by $\alpha p_1, \ldots, \alpha p_K$, αW and αA_0. First, note that doing so in (35) will mean that P will also be replaced by αP. Second, doing so in the budget constraint (26) will multiply each of the terms in it by α. This yields:

$$\Sigma_k \alpha p_k x_k + (R - R_m)\alpha P m = \alpha A_0 + \alpha W n \tag{26'}$$

But canceling out α from both sides of (26′) returns us to (26), so that the first-order conditions (27) to (30) and the solutions given by (31) to (33) for the values of the endogenous variables must be the same for (25) subject to (26′) as for (25) subject to (26). Hence, the quantities demanded of the commodities and real balances and the supply of labor are not affected by a proportionate increase from $(p_1, \ldots, p_K, W, A_0)$ to $(\alpha p_1, \ldots, \alpha p_K, \alpha W, \alpha A_0)$. Formally stated, *the demand and supply functions in (31) to (33) are homogeneous of degree zero* in $p_1, \ldots, p_K, W, A_0$.[29] This property is incorporated in the following set of equations:

$$x_k^{\mathrm{dh}} = x_k^{\mathrm{dh}}(\alpha p_1, \ldots, \alpha p_K, \alpha W, (R - R_m)\alpha P, \alpha A_0) \quad k = 1, \ldots, K \tag{36}$$

$$n^{\mathrm{s}} = n^{\mathrm{s}}(\alpha p_1, \ldots, \alpha p_K, \alpha W, (R - R_m)\alpha P, \alpha A_0) \tag{37}$$

$$m^{\mathrm{dh}} = m^{\mathrm{dh}}(\alpha p_1, \ldots, \alpha p_K, \alpha W, (R - R_m)\alpha P, \alpha A_0) \tag{38}$$

28 Both the CPI and the GDP deflator suffer from certain limitations and imperfections. Descriptions of these can be found in many macroeconomics textbooks.

29 This can also be proved directly from the first-order conditions. To do this, divide (27) and (29) by (28). In the resulting set of equations, replacing $(p_1, \ldots, p_K, W_0$ and $\mathrm{A}_0)$ by $(\alpha p_1, \ldots, \alpha p_K, \alpha W$ and $\alpha A_0)$ does not induce any change. Further, as discussed above in the text, this replacement does not alter (30), which is the budget constraint itself. Hence, the first-order conditions (27) to (30) and their solution (31) to (33) will remain unchanged.

for any $\alpha > 0$. The superscript h in (38) differentiates the household demand for real balances given by (38) from the firm's demand (m^{df}) for them derived later in this chapter.

3.7.4 Relative prices and the numeraire

If we let $\alpha = 1/P$, (31) to (33) yield:

$$x_k^{dh} = x_k^{dh}(p_1/P, \ldots, p_K/P, W/P, (R - R_m), A_0/P) \quad k = 1, \ldots, K \tag{39}$$

$$n^s = n^s(p_1/P, \ldots, p_K/P, W/P, (R - R_m), A_0/P) \tag{40}$$

$$m^{dh} = m^{dh}(p_1/P, \ldots, p_K/P, W/P, (R - R_m), A_0/P) \tag{41}$$

where:

x_k = relative price of the kth commodity
W/P = relative price of labor (real wage rate)
A_0/P = real value of initial endowments.

Equations (39) to (41) assert that the demands for commodities and real balances and the supply of labor depend only upon relative prices – but not on absolute prices – and the real value of initial endowments. These relative prices have been defined in terms of the composite bundle of commodities used in calculating the price level. This composite bundle is here being used as a *numbering device* or *numeraire* for measuring the real cost of the various goods.

In (39) to (41), if we had specified α to equal $1/p_i$ rather than $1/P$, where p_i is the price of a specific good i, good i would have served as the numeraire. In this case, the resulting relative prices p_k/p_i and W/p_i would have been in terms of the numeraire good i rather than of the composite bundle of commodities in the price index.

If we had wanted to express the cost of buying goods in terms of *labor units* – i.e. the hours of work (of the worker with the average wage W) required to buy one unit of a good – we would set $\alpha = 1/W$. Doing so would make the relative price of the kth good p_k/W. Labor would become the numeraire. Many classical economists in the nineteenth and early twentieth centuries, as well as Keynes in *The General Theory*, had used this mode for expressing relative prices. This was partly because the construction, availability and use of the price indices were not common until the 1930s. But it was also partly to allow the traditional classical economists to conduct their analysis of the commodities and labor markets completely in real rather than nominal terms, with monetary factors thereby kept out of their analysis. However, the use of labor as a numeraire for analytical purposes has gone into disuse since the 1940s, and the standard practice now is to express relative prices using the CPI or GDP deflator in the denominator.

3.8 The firm's demand and supply functions for money and other goods

Corresponding to the two ways of introducing real balances into the utility function, real balances can also be introduced directly or indirectly into the production function.

3.8.1 Money in the production function (MIPF)

Assume that the representative firm producing the kth commodity has a production function specified by:

$$x_k = F(n, \kappa, m^{\mathrm{f}}) \tag{42}$$

where:

x_k = quantity of the kth good, $k = 1,\ldots, K$, produced by the firm
n = number of workers
κ = variable physical capital stock
m^{f} = real balances held by the firm.

The rationale for putting the firm's real balances as an input in its production function is that holding them allows the firm to produce greater output with given amounts of labor and capital. If it did not hold any real balances, it would have great difficulty in paying its employees and suppliers or selling its output. To avoid handling payments and receipts in money, the firm would have to divert part of its labor and capital to arrange somehow for payments and receipts directly in commodities, with such diversion reducing the amounts of labor and capital allocated to production and thereby reducing the firm's output. Further, the greater the real balances held by the firm, the easier it is for the firm to handle its payments and receipts and the less the need to divert labor and capital to the exchange processes and away from production. Therefore, in an economy requiring the exchange of goods against money, real balances function as an input in the firm's production function, with higher real balances leading to higher output, so that the marginal productivity of real balances is positive. We will assume that this marginal productivity is diminishing, just as for labor and capital.

Therefore, in (42), the first-order partial derivatives, F_n, F_κ and F_m, are assumed to be all positive, and the second-order ones, F_{nn}, $F_{\kappa\kappa}$, F_{mm}, are all negative. The firm may also have a fixed capital stock, implying that it also has some fixed costs of production.

3.8.2 Money in the indirect production function

It is sometimes argued that money does not directly increase the productive capacity of the firm and should not be put in the production function. However, just as with the indirect utility function, we can specify a production function in which money does not appear directly but does so indirectly. This is done in the following.

We assume that the firm's output depends on its capital and the part of its employment that it uses directly as an input in production. However, it has to divert some of its workers to carrying out transactions involving the purchase of inputs – that is, labor and purchases of raw materials and intermediate inputs – and the sale of its output. In the extreme case where the firm does not hold any balances in a monetary economy, it would have to persuade workers and other input suppliers to accept the commodity it produces as payment. It would also have to pay profits to its owners in the same commodity. If it is a corporation, its distributed profits would have to be in this commodity and, for retained profits diverted to investment, it would have to exchange for investment goods some of the commodity it produces. Any such attempt would prevent the firm from existing in the modern economy. In a less extreme case, if the firm held only a small and relatively inadequate amount of money, it would have to employ workers in juggling its money holdings to carry out the required transactions of purchase

and sale. Holding real balances, therefore, allows the firm to economize on the workers it has to divert to carrying out payments.

These arguments imply the production function to be:

$$x_k = x_k(\kappa, n_1) \tag{43}$$

where both partial derivatives are positive and:

x_k = output of the kth commodity
κ = physical capital stock
n_1 = labor directly involved in production.

Total employment by the firm is n, where $n = (n_1 + n_2)$, so that:

$$n_1 = n - n_2 \tag{44'}$$

where n_2 is the amount of the firm's employment used in making payments. Therefore, $\partial n_1 / \partial n_2 < 0$.

For the labor used in carrying out transactions, and using the firm's output x_k as a proxy for the number of payments involved in purchasing inputs and selling output, the general form of the "payments technology function" for a monetary economy would be:

$$n_2 = n_2(m^{\mathrm{f}}, x_k) \tag{44''}$$

where m^{f} are the real balances held by the firm, $\partial n_2 / \partial m^{\mathrm{f}} \leq 0$ and $\partial n_2 / \partial x_k > 0$. The specific form of $n_2(.)$ would depend on the trading and payments technology of the economy and would shift with that technology. Innovations in the financial system, such as the use of direct deposit of salaries into the workers' accounts, or payments to suppliers by electronic transfers, would reduce the demand for real balances for transactions associated with a given level of output and shift the transactions technology function.

From (43), (44′) and (44″),

$$\frac{\partial x_k}{\partial m^{\mathrm{f}}} = \frac{\partial x_k}{\partial n_1} \frac{\partial n_1}{\partial n_2} \frac{\partial n_2}{\partial m^{\mathrm{f}}} \geq 0$$

A specific form of (44″) is:

$$n_2 / x_k = \Phi(m^{\mathrm{f}} / x_k) \tag{45}$$

where $\phi' = \partial \phi / \partial (m^{\mathrm{f}} / x_k) \leq 0$. For this function, the firm reaches "saturation" in real balances relative to its output when $\phi' = 0$. From (43) to (45),

$$x_k = x_k(\kappa, n - x_k \cdot \phi(m^{\mathrm{f}} / x_k)) \tag{46}$$

which can be rewritten as the indirect production function

$$x_k = f(\kappa, n, m^{\mathrm{f}}) \tag{47}$$

where, as shown earlier, $\partial x_k / \partial m^{\mathrm{f}} \geq 0$. Hence, the use of money by the firm increases its output, with its marginal product being positive up to the saturation point. Up to this point, the usage

of money allows the firm to reduce the labor allocated to transactions, thereby increasing the labor allocated directly to production. This increases the firm's output produced with a given amount of employment.

The preceding analysis provides a rationale for putting money in the production function, even though real balances were not assumed to directly increase output for the firm. Given this result, we will revert in further analysis to the direct production function (42).

3.8.3 Maximization of profits by the firm

The firm is assumed to operate in perfect competition in all (output and input) markets and to maximize profits. Its profits are given by:

$$\Pi = p_k F(n, \kappa, m^{\mathrm{f}}) - Wn - \rho_\kappa \kappa - \rho_m m^{\mathrm{f}} - F_0 \tag{48}$$

where:

Π = profits
ρ_κ = nominal user cost of variable physical capital
F_0 = fixed cost of production.

The nominal user cost ρ_m of real balances was derived above as $(R - R_m)P$. The *user cost of capital* is similarly the rental value of a unit of physical capital (such as a machine) per period. The nominal user cost of physical capital in a perfect market is given by:

$$\rho_\kappa = (R + \delta_\kappa - \pi_\kappa) p_\kappa \tag{49}$$

where:

δ_κ = rate of depreciation of the capital good
π_κ = rate of increase in the price of the capital good
p_κ = price of the capital good.

Since the rate of depreciation of capital does not play any particular role in our further analysis, let $\delta = 0$. Therefore, the nominal user cost of capital in our analysis would be:

$$\rho_\kappa = (R - \pi_\kappa) p_\kappa$$

Hence,

$$\Pi = p_k F(n, \kappa, m^{\mathrm{f}}) - Wn - (R - \pi_\kappa) p_\kappa \kappa - (R - R_m) P m - F_0 \tag{50}$$

The first-order conditions for maximizing profits with respect to n, κ, m^{f}, are:

$$p_k F_n - W = 0 \tag{51}$$

$$p_k F_\kappa - (R - \pi_\kappa) p_\kappa = 0 \tag{52}$$

$$p_k F_m - (R - R_m) P = 0 \tag{53}$$

3.8.4 The firm's demand and supply functions for money and other goods

Dividing each term in the first-order conditions (51) to (53) by the price level P, these conditions become:

$$p_k/P.F_n = W/P \tag{54}$$

$$p_k/P.F_\kappa = (R - \pi_\kappa)(p_\kappa/P)\kappa \tag{55}$$

$$p_k/P.F_m = (R - R_m)m^{\mathrm{f}} \tag{56}$$

Solving the set of equations (54) to (56) yields:

$$n^d = n^d(p_k/P, w, (R - \pi_\kappa)(p_\kappa/P), (R - R_m)) \tag{57}$$

$$\kappa^d = K^d(p_k/P, w, (R - \pi_\kappa)(p_\kappa/P), (R - R_m)) \tag{58}$$

$$m^{\mathrm{df}} = m^{\mathrm{df}}(p_k/P, w, (R - \pi_\kappa)(p_\kappa/P), (R - R_m)) \tag{59}$$

where w is the real wage rate W/P. The superscript d indicates demand and the superscript f indicates the representative firm. To be added to (57) to (59) is the supply function for commodities, obtained by substituting (57) to (59) in the production function (47). This yields:

$$x^{\mathrm{s}} = x^{\mathrm{s}}(p_1/P, \ldots, p_K/P, w, (R - \pi_\kappa)(p_\kappa/P), (R - R_m)) \tag{60}$$

The first-order conditions (54) to (56) imply that (57) to (60) are *homogeneous equations of degree zero* in $p_k, k = 1, \ldots, K, W$ and P. That is, proportionate increases in these variables will not alter the inputs demanded and the output supplied by the firm. Note that this result requires constancy of the user costs $(R - \pi_\kappa)$ and $(R - R_m)$. Note that a change in these would change the demand for inputs and the supply of output.

Since physical capital is a commodity, though both used and produced by firms, the general properties of its demand and supply functions are identical to those of commodities.

3.9 Aggregate demand and supply functions for money and other goods in the economy

Equations (39) to (41) have specified the demand and supply functions of a representative consumer. Aggregating these over all consumers, with the relevant symbols now taken to refer to the respective aggregate, we have, from (39) to (41):

$$x_k^{\mathrm{d}} = x_k^{\mathrm{d}}(p_1/P, \ldots, p_K/P, W/P, (R - R_m), A_0/P) \quad k = 1, \ldots, K \tag{61}$$

$$n^{\mathrm{s}} = n^{\mathrm{s}}(p_1/P, \ldots, p_K/P, W/P, (R - R_m), A_0/P) \tag{62}$$

$$m^{\mathrm{dh}} = m^{\mathrm{dh}}(p_1/P, \ldots, p_K/P, W/P, (R - R_m), A_0/P) \tag{63}$$

where:

x_k^{d} = aggregate demand for the kth commodity
n^{s} = aggregate supply of labor
m^{dh} = households' demand for real balances
A_0 = aggregate initial endowment of all consumers.

Also aggregate (57) to (59) over all firms in the economy, again adopting the convention that the relevant symbols will now refer to the respective aggregates. This operation yields the supply function for commodities and the demand functions for labor and the real balances of firms as:

$$x_k^{\mathrm{s}} = x^{\mathrm{s}}(p_1/P, \ldots, p_K/P, W/P, (R-\pi_\kappa)(p_\kappa/P), (R-R_m)) \quad k = 1, \ldots, K \tag{64}$$

$$x_\kappa^{\mathrm{s}} = x_\kappa^{\mathrm{s}}(p_1/P, \ldots, p_K/P, W/P, (R-\pi_\kappa)(p_\kappa/P), (R-R_m)) \tag{65}$$

$$x_\kappa^{\mathrm{d}} = x_\kappa^{\mathrm{d}}(p_1/P, \ldots, p_K/P, W/P, (R-\pi_\kappa)(p_\kappa/P), (R-R_m)) \tag{66}$$

$$n^{\mathrm{d}} = n^{\mathrm{d}}(p_1/P, \ldots, p_K/P, W/P, (R-\pi_\kappa)(p_\kappa/P), (R-R_m)) \tag{67}$$

$$m^{\mathrm{df}} = m^{\mathrm{df}}(p_1/P, \ldots, p_K/P, W/P, (R-\pi_\kappa)(p_\kappa/P), (R-R_m)) \tag{68}$$

Adding equations (63) and (68) yields the economy's aggregate demand for real balances m^{d} as:

$$m^{\mathrm{d}} = m^{\mathrm{d}}(p_1/P, \ldots, p_K/P, W/P, (r-\pi_\kappa)(p_\kappa/P), (r-r_m), A_0/P) \tag{69}$$

Equations (61) and (64), (65) and (66) are respectively the economy's demand and supply functions for commodities; (65) and (66) are respectively the economy's demand and supply functions for physical capital; (67) and (62) are respectively the demand and supply functions for labor; and (69) is the economy's demand function for real balances. For a complete model of the economy, we are still missing an equation for the supply of real balances to the economy.

3.10 Supply of nominal and real balances

The supply of nominal balances to the economy can be endogenous – that is, a function of some of the other variables in the model – or exogenous. Which of these is the pertinent one to a given economy will depend upon the degree of control that the central bank has over the nominal money supply and whether it considers it preferable to use the money supply or the interest rate as its primary instrument of monetary policy. Until about the mid-1990s, the common assumption in general equilibrium models was that the central bank uses the money supply as its primary instrument of monetary policy and that its amount can be taken as exogenous (see Chapter 10). We adopt this assumption for the following analysis. Designating M as the exogenously supplied money stock, this assumption is that:

$$M^{\mathrm{s}} = M \tag{70}$$

Therefore, the amount of real balances $m^{\mathrm{s}}(= M^{\mathrm{s}}/P)$ supplied to the economy is given by:

$$m^{\mathrm{s}} = M/P$$

Since the price level P is determined endogenously by the model, the supply of real balances m^{s} is an endogenous variable even though the nominal money supply M was assumed to be exogenously determined. That is, while the central bank controls the nominal supply, the economy itself determines the real balances in the economy.

3.11 General equilibrium in the economy

The preceding analysis specifies the equilibrium conditions for all markets as:

The markets for consumer commodities, with $k = 1, \ldots, K$:

$$x_k^{\mathrm{d}}(p_1/P, \ldots, p_K/P, W/P, (R - R_m), A_0/P) = x_k^{\mathrm{s}}(p_1/P, \ldots, p_K/P, w, (R - \pi_\kappa)(p_\kappa/P), (R - R_m)) \tag{71}$$

The market for physical capital:

$$x_\kappa^{\mathrm{d}}(p_1/P, \ldots, p_K/P, W/P, (R - \pi_\kappa)(p_\kappa/P), (R - R_m)) = x_\kappa^{\mathrm{s}}(p_1/P, \ldots, p_K/P, W/P, (R - \pi_\kappa)(p_\kappa/P), (R - R_m)) \tag{72}$$

The labor market:

$$n^{\mathrm{d}}(p_1/P, \ldots, p_K/P, W/P, (R - \pi_\kappa)(p\kappa/P), (R - R_m)) = n^{\mathrm{s}}(p_1/P, \ldots, p_K/P, W/P, (R - R_m), A_0/P) \tag{73}$$

The money market:

$$m^{\mathrm{d}}(p_1/P, \ldots, p_K/P, W/P, (R - \pi_\kappa)(p\kappa/P), (R - R_m), A_0/P) = M^{\mathrm{s}}/P \tag{74}$$

In addition, the definition of the price level from (34) is:

$$P_t = [\Sigma_k \, p_{kt} \, x_{k0}]/[\Sigma_\kappa \, \rho_{k0} \, x_{k0}] \tag{75}$$

(71) to (75) constitute a set of $(K + 4)$ equations in the $(K + 4)$ endogenous variables $p_1, \ldots, p_K$, W, $\rho_\kappa (= R - R_\kappa)$, P and $\rho_m (= R - R_m)$. We follow the usual assumption that since the number of equations equals the number of endogenous variables, a unique solution exists for this system.

The above equilibrium equations (71) to (74) are homogeneous of degree zero in $p_1, \ldots, p_K$, W, A_0 and M^{s}. Therefore, a once-for-all proportionate increase in all prices, and therefore in P, provided the real values of the initial endowments and real balances are held constant, would not change the quantities demanded, supplied and traded in the economy. The real values of the variables would not be affected and neither consumers nor firms would be better or worse off under these changes.

Conversely, *a once-for-all increase in the money supply which results in a once-for-all proportionate increase in all prices* – so that the growth rate of the money supply does not change – will not have any real effects on the economy as long as these increases do not change the real value of initial endowments and do not induce expectations of inflation.[30]

30 It is generally assumed in macroeconomic theory that once-for-all money supply increases do not lead to anticipations of inflation and therefore do not change the rates of interest r and r_m.

The role of initial endowments in general equilibrium analysis

Initial endowments can be in the form of commodities, money or other financial assets (bonds), so that:

$$A_0 = \Sigma_k \, p_k \, \bar{x}_{k,0} + \overline{M}_0 + p_b \, \bar{b}_0 \tag{76}$$

where $\bar{x}_{k,0}$ is the initial endowment of the kth commodity, $\overline{M}_0$ is the carryover of nominal balances and $\bar{b}_0$ is the carryover of real bonds at a market price of p_b. The real value a_0 of endowments is given by:

$$a_0 = A_0/P = \Sigma_\kappa (p_k/P)\bar{x}_{k,0} + \overline{M}_0/P + (p_b/P)\bar{b}_0 \tag{76'}$$

A change in the prices of all commodities does not necessarily imply a proportionate change in the nominal balances carried over from the preceding period or a proportionate change in the price of bonds. If these do not change proportionately, a change in the commodity price level will change the real value of endowments.

Real balance effect

If the money supply is held constant, an increase in P will reduce the initial endowments of real balances, making the individual poorer and causing an income effect on the demands for goods. This income effect, in the normal case, would reduce the demands for commodities and real balances and increase the supply of labor. The name given to this effect of changes in the real money stock on the aggregate demand for commodities, and other goods, is the *real balance effect*. Note that it can occur through a change in the price level or in the money supply, but it does not come into play if both the money supply and the price level change in the same proportion.

The real balance effect is an important analytical mechanism connecting the commodity sector to the monetary one (Patinkin, 1965). To illustrate, suppose that the money supply increases. Until prices change, this increase in the money supply increases the real value of real balances and, therefore, of endowments. This will increase the demand for commodities, creating an excess demand in the commodity markets and pushing up their prices. The real balance effect, therefore, provides a mechanism by which changes in the money supply bring about changes in the price level.

Alternatively, suppose the economy is in general equilibrium. A shock that reduces the aggregate demand for commodities will lower the price level and might also raise unemployment. But this price decrease will increase real balances, which, in turn, will serve to increase the demand for commodities. This increase in commodity demand will continue until real balances return to their original equilibrium level. This will require that the price level return to its original level. Hence, the real balance effect functions as an equilibrating mechanism and a link between the monetary and the commodity markets. As such, it rejects any assertions about the dichotomy, discussed later in this chapter, between the real and the monetary sectors.

However, empirically the real balance effect is of little practical significance as a determinant of consumption, so that the absence of a dichotomy between the real monetary

sectors has to rely on some other basis. The real balance effect and its related Pigou effect are more extensively presented in Chapters 14 and 18.

Market for bonds and the interest rate

Initial endowments also include all non-monetary assets which we have termed "bonds." The relationship of the prices of bonds (including equities) with the commodity price level and the inflation rate is still not well understood in macroeconomics. The usual assumption is that their real value is homogeneous of degree zero in the price level P. However, this is more of a convenient assumption rather than one whose validity is generally accepted. Consequently, besides the real balance effect, there may also be a "bonds effect" – that is, an income effect from changes in the real value of bonds (e.g. induced by changes in the price level) on the demand for commodities.

The preceding general equilibrium analysis does not incorporate the market for bonds. Bonds, which were assumed to be illiquid, are a mechanism for transferring purchasing power from the present to the future. Their proper analysis requires an intertemporal framework, so that the preceding one-period analysis is unsuitable for the analysis of the demand for and supply of bonds. Since the return on bonds is the nominal interest rate R, this rate is not determined in the above static model and has to be taken to be exogenously specified, as is the quantity of bonds traded in the economy. However, the above model does determine the user costs of physical capital and real balances.

Further consideration of the market for bonds in the macroeconomic context is presented in Chapters 17 and 20.

3.12 Neutrality and super-neutrality of money

3.12.1 Neutrality of money

The *neutrality of money* is said to exist if *once-for-all* changes in the money supply do not affect the real values of the variables – such as output, employment consumption, real wages, real interest rate and even real balances – in the economy. Another way of expressing this neutrality is to say that money is a *veil*: while its presence – as against its absence in a barter economy – makes a vital difference, changes in it do not have any real effects. The preceding section proves the neutrality of changes in the money supply *in general equilibrium* if:

1 All prices increase in the same proportion.
2 The real value of the initial endowments does not change.
3 Interest is paid on *all* money balances.
4 There is no anticipation of further price changes.

Hence, under these conditions, a once-for-all increase in the money supply, no matter how large, can be ignored for all real purposes since it would have no real effects.

3.12.2 Super-neutrality of money

The *super-neutrality of money* is said to exist if continuous changes in the money supply do not have any real effects.

Continuous increases in the money supply usually result in continuous inflation and such inflation is bound to be wholly or mostly expected. Lenders want the rates of interest to rise by the expected rate of inflation, so as to compensate them for the loss through inflation of the purchasing power of the funds that are lent. Therefore, in perfect money and capital markets, the (Fisher) relationship (approximate for low values of r and π^e) between the interest rates and expected inflation is:

$$R = r + \pi^e \tag{77}$$

$$R_m = r_m + \pi^e \tag{78}$$

where:

π = rate of inflation
π^e = expected rate of inflation
r = real rate of interest (paid by bonds)
r_m = real rate of interest on real balances

and

$$(R - R_m) = (r - r_m)$$

so that even continuous anticipated inflation does not affect the real user cost of real balances.

Assuming $\pi^e = \pi$ (that is, inflation is fully anticipated) for a period of continuous *systematic* inflation,[31] we have:

$$R = r + \pi$$

$$R_m = r_m + \pi$$

Further, since capital goods are also commodities, assume that the inflation in the capital goods price is the same as on all the other goods in the economy, so that $\pi_\kappa = \pi$. This implies that $(r - \pi_\kappa)$ can be replaced by r^r in all relevant equations. Under these assumptions, $(R - R_m)$ can be replaced by $(r - r_m)$ and $(R - \pi_\kappa)$ can be replaced by r in (71) to (75).

Consequently, *if the nominal value of the initial endowments increases by the rate of inflation*, the identical rates of inflation in all prices (including wages) will not change the general equilibrium solution. Therefore, continuous money supply increases, which induce continuous inflation and simultaneously change the nominal value of the initial endowments by the rate of inflation, would not change the demands and supplies in the economy and therefore would not change output, employment, the real rate of interest and real balances. Hence, the super-neutrality of money will hold in general equilibrium under the assumptions:

1 All prices increase in the same proportion.
2 The real value of the initial endowments does not change.
3 Interest rate R_m is paid on *all* money balances.
4 The expected inflation rate equals the actual rate, so that there are no errors in inflationary expectations.

31 This follows from rational expectations. The theory of rational expectations is presented in Chapter 8.

3.12.3 Reasons for deviations from neutrality and super-neutrality

Among the reasons for the non-neutrality of money are:

- Some components, e.g. currency and most forms of checking deposits, do not pay interest. For such components, $R_m = 0$, which affects their demand. Further, changes in the inflation rate will change the cost of using money and therefore change its demand. These changes will change the solution to the set of equations (71) to (75), so that the real output, employment, the real rate of interest and the real values of the other endogenous variables will be altered. Hence, if some or all of the components of the money supply do not pay interest, the neutrality and super-neutrality of money and inflation will no longer apply.
- The neutrality of money requires that the real value of initial endowments does not change. But this value tends to change in disequilibrium. Whether the increases in the money supply and inflation change the real value of endowments or not will depend upon how money is introduced into the economy and the structure of the economy. If the money supply is introduced through open market operations, the increase in the money supply will be counterbalanced by the decrease in the nominal value of the bonds in the hands of the public, so that the nominal value of the initial endowments (which include both bonds and nominal balances) will remain unchanged while their real value will fall. This implies that the super-neutrality of money will not hold.
- The constancy of the real value of initial endowments requires that the ratio of bond prices to the price level (of commodities) remains invariant to changes in the money supply and the other economic adjustments to it. Note that the term "bonds" covers all non-financial assets so that "bond prices" include stock market prices. Economics has no generally accepted theory that the required invariance of the relative prices of bonds and equities to money supply changes does hold. In fact, it is highly plausible, on the basis of everyday experience, that it does not hold for the impact period and the short run. Further, we also need the invariance, of the prices of physical capital and durable consumer goods, including housing, relative to the price level to ensure the neutrality of money. This is also highly questionable for the impact period and the short run. Hence, the invariance of the real value of initial endowments – the wealth of the economy – to money supply changes is highly doubtful in the short run. It may hold for the long run.
- Prices, incomes or wages may be sticky or rigid for some time. For instance, prices are costly to change on a continuing basis so that certain delays in changing them are profit maximizing. Nominal wages are fixed for the duration of labor contracts.
- Many types of income such as pensions, social security payments, unemployment insurance benefits, etc., are changed at infrequent intervals or are not changed sufficiently to match the inflation rate.
- In addition, in economies with pervasive uncertainty, especially about the values of variables – such as the rate of return on investment – influenced by events far in the future, the expected real values of the variables may not be invariant to the rate of inflation.

Some of these topics are discussed at greater length in Chapter 15 on Keynesian economics.

Monetary non-neutrality in disequilibrium

In the adjustment or disequilibrium phase in which the money supply increase has not yet resulted in equi-proportionate increases in the absolute prices of all the commodities or

in the nominal wage rates, the relative prices of commodities and the real value of initial endowments would change, causing real changes in the economy. Hence, money is not neutral in the disequilibrium state of the economy. On a practical note, it is difficult to determine whether disequilibrium is a transitory state, with rapid adjustment to equilibrium, so that its consequences are minimal and can be ignored. The modern classical school assumes that the economy tends to equilibrium rapidly enough to allow one to focus on equilibrium states only. The Keynesian school believes that the economy can persist in less than full employment disequilibrium for long periods, so that the disequilibrium phases cannot be ignored and may well be designated as under-employment equilibria. Money is not neutral in these states.

In analyses of disequilibrium and the business cycle, the nineteenth-century classical economists had argued that capital goods prices and consumer goods prices did not always change in the same proportion. To illustrate their ideas, consider Wicksell's analysis of the effects of a money supply increase in the pure credit economy, as presented in Chapter 2 above. Suppose the banks lower the market rate of interest. This makes it profitable for the firms to increase their borrowings from the banks for the purpose of increasing their investment. The increase in investment increases the demand for capital goods and increases their price, but there is yet no effect on consumer goods prices. That is, in this phase, p_K/P increases. Further, the increased production of capital goods would require increased employment in this industry, changing the structure of output and employment between the consumer and capital goods industries. Once the increase in investment has been accomplished and workers are spending their increased earnings, consumer goods prices will rise, so that in the later phases of the fluctuation p_K/P will fall back to its equilibrium value. Hence, fluctuations in p_K/P are a fundamental part of the adjustment process by which money affects the economy, and these fluctuations cause fluctuations in the output of different industries and overall employment. Such an analysis was not confined to Wicksell but was a part of traditional classical economics generally, and played an important role in the nineteenth and early twentieth century studies of the business cycle. It disappeared from macroeconomics based on the IS–LM models since such models do not distinguish between the consumer and capital goods industries.

Overall assessment of the departures from the neutrality and super-neutrality of money

The preceding arguments provide a very extensive list of reasons why monetary neutrality may not hold. Hence, at least in the short run, money is not likely to be neutral in the disequilibrium and even the equilibrium phases of an economy in which there is a once-for-all increase in the money supply. It is even less likely to be neutral if there are continuous and variable increases in the money supply.[32]

Therefore, on a practical basis in real world economies, increases in the money supply and inflation do have real effects. It is, however, difficult to determine whether these departures from neutrality are relatively unimportant and transient – as the neoclassical and modern classical schools claim – or very important – as the Keynesian school claims. What does seem to happen is that any economy with persistently high rates of anticipated inflation does adjust its contractual and institutional arrangements to minimize the impact of inflation

32 We will return to this topic again in the macroeconomics chapters 13 to 17.

on the real variables – including the relative prices of commodities, real wages, etc. – in the economy, so that the departures from the neutrality of money are reduced. The larger departures from neutrality of money occur when a significant part of the inflation rate is unanticipated. This tends to occur in periods of fluctuating money supply growth rates and inflation rates.

3.13 Dichotomy between the real and the monetary sectors

The *neutrality of money* in general equilibrium is, as shown above, related to the homogeneity of degree zero of all demand and supply functions with respect to changes in all absolute prices and initial endowments. The traditional classical economists sometimes extended their arguments to assert the dichotomy between the real and the monetary sectors. This dichotomy is the statement that the real values of the endogenous variables in the economy are independent of the nominal money supply and demand, and of the price level, so that these real values of the variables can be determined independently of the latter factors.

We can define the *weak* form of the dichotomy as one where the preceding statement holds only in equilibrium and the *strong* form as one where that statement holds both in equilibrium and in disequilibrium. The following modifies the Walrasian general equilibrium set of equations to produce the strong form of the dichotomy between the real sector and the monetary one.

Strong dichotomy and the independence of the real sector from the monetary sector

The general equilibrium system of equations (71) to (74) for the real sector of the economy has the money stock as a component of the initial endowments, as well as the price level as a variable. Modifying these equations to show the complete independence of the real variables from the money supply and the price level requires exclusion of the financial part of the endowments from these equations and the elimination of the price level. For the former, assume that the endowments consist only of commodities. That is, rather than using (76) to describe the initial endowments, the assumption now is:

$$A_0 = \Sigma_k p_k x_{k,0} \tag{79}$$

where $x_{k,0}$ is the initial endowment of the kth commodity, so that there are no carryover money balances or bonds in this system.

To eliminate the price level from the relevant equations, let the numeraire be the first commodity, so that all prices in the economy will be measured in terms of this commodity. This means dividing all absolute prices by p_1 rather than P, which is permitted by the homogeneity of degree zero of the equilibrium equations (71) to (74). Before doing so, restate (79) as:

$$a_0 = A_0/p_1 = \Sigma_k (p_k/p_1) x_{k,0} \tag{79'}$$

Using the first commodity as a numeraire means that $p_1 = 1$, so that p_1 (the price of the numeraire commodity in terms of money) cannot be determined in this model. We correspondingly omit the equilibrium condition for the numeraire commodity by virtue of Walras's law (see Chapter 18), which implies that if equilibrium exists in all markets except

one it must also exist in the remaining market. We also delete the user cost of real balances $\rho_m(= R - R_m)$ as a variable from our system, for reasons of consistency with the historical debates on this topic which excluded any variables related to money from the specified system. The resulting system is:

Modified commodities markets equations for k = 2, …, K:

$$x_k^{\mathrm{d}}(p_2/p_1, \ldots, p_K/p_1, W/p_1, a_0) = c^{\mathrm{s}}(p_2/p_1, \ldots, p_K/p_1, W/p_1, \rho_k/p_1) \tag{80}$$

Modified physical capital market equation:

$$x_\kappa^{\mathrm{d}}(p_2/p_1, \ldots, p_K/p_1, W/p_1, \rho_k/p_1) = x_\kappa^{\mathrm{s}}(p_2/p_1, \ldots, p_K/p_1, W/p_1, \rho_k/p_1) \tag{81}$$

Modified labor market equation:

$$n^{\mathrm{d}}(p_2/p_1, \ldots, p_K/p_1, W/p_1, \rho_k/p_1) = n^{\mathrm{s}}(p_2/p_1, \ldots, p_K/p_1, W/p_1, a_0) \tag{82}$$

Note that (80) to (82) assume that there are no carryover money balances or bonds, so that these equations are not valid for an economy in which such carryover between periods exists, as it does in all real-world economies. These are $(K + 1)$ equations in the $(K + 1)$ endogenous real variables $(p_2/p_1, \ldots, p_\kappa/p_1, W/p_1, \rho_k/p_1)$. As in earlier analysis, since this is a one-period and not an intertemporal analysis, the interest rate R on bonds is taken as given to this model. The money supply and the price level are not in these equations. Assuming that a solution exists, (80) to (82) can be solved for the real values of the endogenous variables, even without knowing the amount of nominal balances in the economy or knowing the price level. Hence these equations represent a strong form of the dichotomy; in it, the real sector by itself determines its relative prices, quantities demanded and supplied, and output of commodities, as well as employment in the economy. This determination is independent of the economy's money supply or the price level, so that changes in them cannot affect the real sector of the economy.

Strong dichotomy and the determination of the price level

The relative prices determined from (80) to (82) could be substituted in the money market equation, which – with a_0 replacing A_0/p_1 and omitting ρ_m – would be given by:

Modified money market equation:

$$m^{\mathrm{d}}(p_1/p_1, \ldots, p_K/p_1, W/p_1, \rho_k/p_1, a_0) = M^{\mathrm{s}}/P \tag{83}$$

Since the real values of all the arguments of the m^{d} function on the left side are determined by equations (80) to (82) independently of the money supply and the price level, (83) can be rewritten as:

$$P = [1/m^{\mathrm{d}}(.)]M^{\mathrm{s}} \tag{84}$$

Equation (84) determines the price level, so that the money market equilibrium is needed for the determination of the price level but not for the determination of the values of the real variables.

The traditional classical economists had, in fact, used the quantity theory of money instead of the more general money demand function in (83). Therefore, their version of (83) was:

$$M^{s} = m_{y}Y = m_{y}Py^{f}$$

where y^{f} was pre-determined by the real sector of the economy and M^{s} was exogenous, so that the quantity theory determined the price level for the economy.

Equation (84) implies that the price level will change in proportion to the money supply. This was the central proposition of the quantity theory, so that the quantity theory and the traditional classical notion of the dichotomy between the real and monetary sectors were consistent with each other and supported each other in economists' thinking.[33]

Strong dichotomy and the determination of the velocity of money

Note that, for this dichotomous Walrasian system, the velocity v of circulation of money is given by:

$$\begin{aligned} v &= Y/M = y/m \\ &= y/m^{d}(.) \end{aligned} \tag{85}$$

Since the dichotomized Walrasian system determines both y and m^{d} as real variables, independently of the money supply and the price level, the equilibrium velocity of circulation in this system is also a real variable and is independent of the money supply and the price level, as Irving Fisher had claimed (see Chapter 2 above).

Strong dichotomy and the indeterminacy of the price level

It is important to note that a strongly dichotomized system produces indeterminacy of prices: a change in the demands and/or supplies of commodities does not compel the markets for these commodities to change the absolute prices of these commodities, since absolute prices are not variables in these functions. Conversely, an arbitrary change in the price level does not change any demands or supplies in real terms and does not change their equilibrium solutions. Therefore, any arbitrarily specified price level is consistent with the real sector of such an economy. Further, an increase in the money supply will not increase the demand for individual commodities or the aggregate demand for commodities and, therefore, will not put pressure on the individual prices and the price level to rise, so that we are left without a mechanism for price increases.

Real balance effect and overall assessment of the strong dichotomy

Since money balances by their nature act as a store of value that must be carried over from one period to the next, they must form part of the initial endowments of the individuals in

33 However, neither alone was sufficient to imply the other.

the economy and of the whole economy. That is, it is not legitimate to rewrite the set of equilibrium conditions (71) to (73) as the set (80) to (82). Hence, monetary economies do not possess dichotomy between the real and the monetary sectors, and the conclusions based on the dichotomy have to be rejected.

The critical element in the link from the monetary to the real sector is the real balance effect. This link operates in the disequilibrium phase, and does so through changes in the real balances impacting on the demand for commodities. It is primarily a mechanism operating in the disequilibrium phase of the Walrasian system and will not be noticeable in the static equilibrium description of such a system, so that the latter may seem to indicate the dichotomy even when one does not really exist. We will return to this issue again in Chapter 18 on Walras's law and the interactions among the sectors of the economy.

3.14 Welfare cost of inflation

The welfare costs of inflation can arise from several sources. Among these are:

1 impact of inflation on money demand;
2 seigniorage to the government from inflation;
3 impact of inflation on output and unemployment;
4 impact of inflation on the informativeness of relative prices for economic decisions;
5 welfare costs of inflation due to the rigidity of nominal payments in contracts.

The following subsections discuss these. As they indicate, the assessment of the overall net cost of inflation is difficult to calculate. However, most of the above categories, except possibly for (3), impose welfare losses. It is generally accepted that inflation above a very small percentage, consistent with price stability in an environment with improvements in product quality and the introduction of new products, imposes net welfare losses, so that it is preferable for policy makers to aim for an inflation rate basically consistent with price stability. Driffill *et al.* (1990) and Gillman (1995) provide surveys of the costs of inflation.

Welfare cost of inflation from its impact on money demand

Our analysis in this and preceding chapters shows that money demand is negatively related to the nominal interest rate, and the Fisher equation on interest rates shows that, with perfect capital markets, the nominal interest rate rises by the expected rate of inflation. Hence, expected inflation decreases holdings of real balances. Therefore, the demand for real balances plotted against the inflation rate will have a downward sloping curve, where the inflation rate is part of the opportunity cost, which is the nominal interest rate, of holding money. Intuitively, as Chapter 4 shows, making do with lower money holdings requires more trips to the bank to convert non-monetary financial assets into money, implying greater inconvenience and effort for the individual.

The analysis of this chapter has been based on real balances being in the utility function, so that smaller holdings of real balances imply less utility than larger ones. This implies that, if there is no cost of creating money and if money holdings do not pay interest, the area under the demand curve for real balances can be used as a measure of the consumer surplus lost as interest rates rise. For this analysis, the demand curve for real balances is drawn against the nominal interest rate. Figure 3.1 shows the demand curve for real balances m^{d}. For an economy in which money is neutral (so that it does not change output) and assuming that

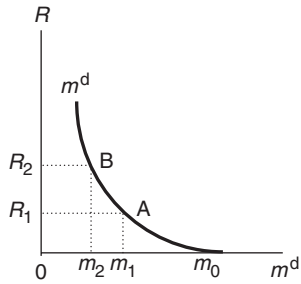


Figure 3.1

money balances do not pay interest, Bailey (1956) measured the welfare cost of holding smaller money balances at the nominal interest rate R_1 than at a zero interest rate by the area $0R_1Am_0$ under the demand curve m^d, since this area measures the consumer surplus lost as a result of a positive nominal interest rate. The welfare cost of inflation at the rate π_1, which causes the nominal interest rate to rise to R_2, with a consequent fall in money demand to m_1, is measured by the area R_1R_2BA under the money demand curve. The estimates of such a cost of inflation differ considerably. Gillman (1995) surveys the welfare losses/costs of inflation and comes to the conclusion that this cost for the USA ranges from 0.85 percent to 3 percent of GNP for every percentage increase in the inflation rate.

The welfare losses from inflation really arise from the interest rate on bonds less that, if any, on money. For this purpose, the demand curve used for measuring the welfare costs of holding smaller balances can be plotted against the bond interest rate less the one on money. If this interest rate differential were zero, there would be no such losses. One way of driving this difference to zero is by inducing an anticipated rate of deflation equal to the real interest rate. However, even anticipated deflation does impose various other costs on the economy. The other way would be to pay interest on money balances equal to that on bonds, which would be difficult and costly to do on currency holdings, though it would not be difficult to arrange on inside money (i.e. deposits in banks).

Walsh (2003, Ch. 2) provides an overall survey of the analytic and empirical findings of the welfare costs of inflation.

Seigniorage to the government from inflation

Compared with bond-financed government expenditures, the government either directly or through the central bank receives revenue from money creation, which allows it to buy resources from the private sector without having to pay interest on bonds and repay the principal borrowed when the bond is retired. Under the simplifying assumption that the rate of inflation equals the rate of increase in the money supply, the government revenue from inflation π_1 and real balances m_1 would be $\pi_1 m_1$. This seigniorage[34] (i.e. revenue from

34 Estimates of seigniorage as a percentage of GDP usually place it at less than 2 percent for most countries, though for countries with high inflation the percentage has been estimated to be as high as 10 percent or so. However, seigniorage as a percentage of government spending is much higher. While it is usually estimated at less than 10 percent, estimates for some countries are 20 percent or more (Click, 1998).

money creation) will reduce the need for tax revenues by a corresponding amount. Tax revenues of the kind usually imposed on the economy imply their own distortions and welfare costs. Some economists claim that the welfare costs of inflation exceed the costs from taxation. However, this conclusion is more likely to apply to developed economies with well-developed and low-cost tax collection systems. It may not apply to the same extent to poor developing countries, which find it difficult and costly to collect adequate tax revenues.

Chapter 22 provides further discussion of seigniorage from inflation as a tax revenue device in the context of an overlapping generations model.

Impact of inflation on output and unemployment

In the short run, inflation has an impact on unemployment. This impact is often captured by some form of the Phillips curve, which has a convex downward slope between inflation on the vertical axis and unemployment on the horizontal one. The forms of the Phillips curve explored in this book are the original Phillips curve (see Chapter 15), the expectations-augmented Phillips curve (see Chapter 14) and the new Keynesian Phillips curve (see Chapter 15). There are, however, disputes about which one of these is valid in the short run. If inflation does reduce short-run unemployment it would increase output, which would constitute a gain from inflation. However, this gain would be short lived and needs to be adjusted by a loss in output from subsequent disinflation.

For the long run, most macroeconomic theories assert that output and unemployment are independent of inflation. However, hyperinflation (i.e. very high inflation) does lead to both short-term and long-term dislocations of the economy and is known to reduce output and increase unemployment severely.

Impact of inflation on the informativeness of relative prices

In market economies, relative prices of commodities play a very useful role in guiding decisions on consumption and production. Since inflation usually does not always increase all prices in the same proportion, it produces changes in relative prices, which can lead to costly mistakes in purchases and production. In labor markets, changes in relative prices of products can lead to different rates of increase in nominal wages among industries and firms and cause industrial unrest, producing increases in strikes. These would mean a misallocation of the economy's resources in production and would be a component of the welfare costs of inflation.

Welfare costs of inflation due to rigidity of nominal payments in contracts

Inflation causes errors in its anticipation, so that it always has an unanticipated component. Since this component of inflation is unexpected, it cannot be accurately incorporated in contracts involving future payments set in nominal terms. Those making payments benefit from an unanticipated increase in inflation, while those receiving payments lose. While this may be classified as a distribution effect, it can have real effects on consumption, production and investment. Further, certain types of contracts set payments in nominal terms at the current price level and do not incorporate future increases in payments to compensate for expected inflation. While indexation to inflation can, in principle, incorporate compensation for inflation, inflation-indexation is not usual.

Certain types of contracts are of extremely long duration. Among these are pensions, mortgages, long-term bonds, etc., so that the impact of anticipated and unanticipated inflation accumulating over time can persist for long periods – and create winners and losers over long periods. The government is often among the beneficiaries of inflation since it collects taxes at set rates on nominal incomes, which rise by the inflation rate. It also pays pensions, which are usually not indexed fully to inflation, so that real pensions decline. It also has a large outstanding amount of long-term nominal bonds, which have a commitment to making payments at nominal coupon rates, so that the real value of coupon payments falls with inflation.

Conclusions

This chapter has provided a basis for both the microeconomic analysis of the demand for and supply of money and the macroeconomic analysis of the role of money in the macroeconomy. The former is further developed in Chapters 4 to 10 and the latter in Chapters 13 to 17.

The analysis of this chapter is in the tradition of the MIUF and MIPF models. This approach treats money as a good like other goods in the utility function and as an input like other inputs in the production function. It puts real balances in the utility function since, for a given individual in a monetary economy, more real balances are preferred to less. A more distinctive approach that would keep them out of both the direct and indirect utility and production functions is offered by the overlapping generations models presented in Chapters 21 to 23.

Money is neutral in the neoclassical model derived in this chapter for a once-only increase in prices, provided that:

- There is a proportionate increase in all prices (including wages).
- The real value of initial endowments remains unchanged.
- The expected rate of inflation remains unchanged.

In this model, money is also superneutral for continuous increases in the price level, provided that:

- All prices rise in the same proportion.
- The real endowments do not change.
- The expected inflation rate is identically equal to the actual rate of inflation.
- The nominal rates of return on bonds, physical capital and *money* all increase by the rate of inflation.

These are fairly stringent conditions. Whether the deviations from neutrality and super-neutrality for a given real-world economy are significant or not would depend upon the particular characteristics of the economy.

Modern classical economists tend towards acceptance of neutrality of money, and sometimes even of super-neutrality, as an acceptable though rough approximation to reality. Keynesian economists tend to consider these as poor and unacceptable approximations and believe that money is not neutral in real-world economies. Their reasons for this are discussed in Chapter 15 and include their belief that the commodity and labor markets do not clear

fast enough.[35] This discussion and its implications for monetary policy are pursued further in the macroeconomics chapters 13 to 17.

The property of neutrality is different from that of dichotomy between the real and the monetary sectors. The strong form of the latter makes the real sector independent of the monetary sector even in disequilibrium, so that changes in the money supply do not affect relative prices and employment. The strong form of the dichotomy therefore does not apply in monetary economies. The link from the monetary sector to the real one is the real balance effect. The link from the financial sector as a whole to the real one is the wealth effect operating through changes in the real value of bonds and money balances. These concepts, as well as neutrality and dichotomy, are discussed in greater detail in Chapter 18 on Walras's law, Say's law and the interrelationship between sectors.

While empirical evidence supports the neutrality of money for output over long periods, it also shows persistent effects of monetary policy on output and unemployment over periods of a few years, or over the business cycle. Chapters 1 and 14 provide the stylized facts on these effects. Most economists now believe that the short-run observations on the non-neutrality of money cannot be explained by Walrasian models with perfect competition and perfect information. They attribute the short-run real effects to market imperfections, staggered overlapping nominal wage contracts and adjustment costs of various types. These issues and models are presented in Chapter 15.

The stylized facts of the relationship between money and output specified in Chapters 1 and 14 also make the point that increases in the money supply initially increase output and only with a lag adjust prices or inflation, so that much of the impact of the money supply does not proceed through the prior adjustment of prices by markets. This advocates caution in the use of the general equilibrium model presented in this chapter. Therefore, the contributions of the general equilibrium model and its implications really belong to the long run, rather than to the short run or business cycle fluctuations.

This chapter has also derived the general demand function for real balances as part of the Walrasian system and examined its properties. The following three chapters use Keynes's motives for holding money to present further analytical developments specific to each motive.

Summary of critical conclusions

- Money can be an argument of the individual's utility function and the firm's production function, either directly or indirectly.
- The user cost of money balances as a medium of payments is not the price level but their rental cost and is represented by the interest foregone from holding them.
- The demand for real balances and the demand and supply functions for all other goods are homogeneous of degree zero in all prices *and* (the nominal value of) each individual's

35 There is an even more fundamental basis for divergence between the neoclassical and Keynesian schools on these issues. Neoclassical economics assumes that all markets tend to clear rapidly so that the analysis can be conducted under the assumption of equilibrium in all markets. This assumption is at the core of the demand and supply functions derived in this chapter. The proper name for such functions is "notional." The alternative to such functions are the "effective" or "quantity-constrained" demand and supply functions which do not assume market clearance in other markets. Such functions have not been derived or presented in this chapter, they belong in the Keynesian tradition and some discussion of them is presented in Chapter 15 on Keynesian economics.

initial endowments. The latter requires homogeneity of degree one of the nominal value of initial endowments in all prices.

- ❖ Omitting initial endowments from the demand and supply functions creates a dichotomy between the real and monetary sectors of the Walrasian general equilibrium model.
- ❖ Keeping initial endowments in the analysis introduces the real balance and wealth effects as a connecting link from the monetary sector to the real one, which implies that the financial and real sectors are intertwined in disequilibrium.
- ❖ Money is neutral – and could be even super-neutral – under rather strict assumptions that do not hold in practice. In particular, changes in all prices are usually not accompanied by a proportionate increase in the nominal value of initial endowments.
- ❖ Among the components of initial endowments that, in practice, tend not to be homogeneous of degree one in all prices are money balances, minimum wage rates, pensions, and bond and equity prices.
- ❖ In the short run, money is not neutral in the economy, but the real question for monetary economics is not a black or white one but rather the degree and duration of the deviations from neutrality for the economy and the period in question.

Review and discussion questions

1. Define the neutrality of money.
   Provide at least a rough proof that money is neutral in a Walrasian general equilibrium.
   Can disequilibrium occur in the Walrasian model? If it can, would money neutrality also exist in disequilibrium in this model? If not, why is money neutrality usually identified with the Walrasian model?
2. For the Walrasian model, discuss the statement: if nominal wages and prices are fully flexible, then neither a one-time increase in the money supply nor an increase in the rate of monetary growth will have any effect on the level of output in general equilibrium.
3. Discuss the relationship in Walrasian general equilibrium analysis between the neutrality and super-neutrality of money and the classical dichotomy. Does either of them imply the other?
4. Discuss the relationship between the neutrality (and super-neutrality) of money and the quantity theory of money. Does either of them imply the other?
5. How important are deviations from the neutrality of money likely to be at single-digit but constant rates of inflation? How important are deviations from the neutrality of money likely to be at single-digit but variable and highly uncertain rates of inflation?
6. Does the neutrality of money hold in hyperinflations? Discuss.
7. Discuss: if all prices, including nominal wages, are flexible, money must be neutral.
8. Why is so much attention paid to initial endowments in the individual's utility analysis? Suppose that initial endowments were left out of such analysis. What analytical consequences would this imply for neutrality and dichotomy, and for the role of money in the macroeconomy?
9. Assume that the representative individual has the specific utility function:

$$U(c, n, m^h) = U(c + m^h - h(n))$$

where c is the purchase of commodities, n is the supply of labor, m^{h} is real balances held by the individual, and $h(n)$ represents the dislike for work or the loss of leisure

due to labor supplied, with $\partial h/\partial n > 0$ and $\partial^2 h/\partial n^2 < 0$. Also assume that, each period, he receives an exogenously specified pension in nominal terms and also earns labor income from his labor supply. Derive the relevant demand and supply functions for the individual, stating any assumptions that you need to make. Are these functions invariant with respect to a proportionate increase in all prices? If not, what is required to make them invariant?

10. Assume that the specific production function of the representative firm is:

$$F(K, L, m^{\mathrm{f}}) = AK^{\alpha}L^{\beta}m^{f\,\gamma}$$

where $F(.)$ is the firm's production function, K is its capital stock, L is its employment and m is its holdings of real balances. Derive the relevant demand and supply functions for the firm, stating any assumptions that you need to make. Define the user cost of money. Show the dependence of marginal productivities of labor and capital on the user cost of real balances.

Suppose that a financial innovation multiplies the firm's marginal productivity of real balances by λ. What would be its impact on the firm's demand functions for labor and capital?

11. "Putting money into the utility and production functions is difficult to reconcile with the Walrasian general equilibrium model." "Putting money in the utility and production functions does provide a way of theorizing about the benefits from the medium of payments role that money plays in the real-world economy with heterogeneous goods, specialization in production and trade, and absence of the double coincidence of wants." Discuss these statements.

References

Bailey, M.J. "The welfare costs of inflationary finance." *Journal of Political Economy*, 64, 1956, pp. 93–110.

Burstein, M.L. *Modern Monetary Theory*. London: St Martin's Press, 1986.

Click, R.W. "Seigniorage in a cross-section of countries." *Journal of Money, Credit and Banking*, 30, 1998, pp. 154–71.

Driffill, J., Mizon, G.E. and Ulph, A. "Costs of inflation." In B. Friedman and F. Hahn, eds, *The Handbook of Monetary Economics*, Vol II. New York: North-Holland, 1990, pp. 1012–66.

Friedman, M. "The quantity theory of money – a restatement." In M. Friedman, ed., *Studies in the Quantity Theory of Money*. Chicago: Chicago University Press, 1956, pp. 3–21.

Gillman, M. "Comparing partial and general equilibrium estimates of the welfare costs of inflation." *Contemporary Economic Policy*, 13, 1995, pp. 60–71.

Holman, J.A. "GMM estimation of money-in-the-utility-function model: the implications of functional forms." *Journal of Money, Credit and Banking*, 30, 1998, pp. 679–98.

Patinkin, D. *Money, Interest and Prices*, 2nd edn. New York: Harper & Row, 1965, Chs. 2, 5–8.

Walsh, C.E. *Monetary Theory and Policy*. Cambridge, MA: MIT Press, 2003.

Part III

The demand for money

4 The transactions demand for money

Keynes had designated the transactions demand for money as due to the transactions motive but had not provided a theory for its determination. In particular, he had assumed that this demand depended linearly on current income but did not depend on interest rates.

Subsequent contributions by Baumol and Tobin in the 1950s established the theory of the transactions demand for money. These contributions showed that this demand depends not only on income but also on the interest rate on bonds. Further, there are economies of scale in money holdings.

The transactions demand for money is derived under the assumption of certainty of the yields on bonds, as well as of the amounts and time patterns of income and expenditures.

Key concepts introduced in this chapter

- Transactions demand for money
- Economies of scale in money demand
- Elasticity of the demand for real balances with respect to the price level
- Elasticity of the demand for real balances with respect to income
- Elasticity of the demand for real balances with respect to the interest rate
- Elasticity of the demand for real balances with respect to their user cost
- Efficient funds management

This chapter presents the main elements of the theory of the demand for transactions balances. In doing so, it follows Keynes in assuming that an individual's money holdings can be validly subdivided into several components, one of which is purely for meeting transactions.

Chapter 2 has pointed out that many of the classical economists and Keynes had made the simple assumption of the unit elasticity of demand for transactions balances with respect to nominal income. In particular, the demand for transactions balances was taken to double if either the price level or real income/expenditures – but not both – doubled. Hardly any analysis was presented in support of such a statement and it remained very much in the realm of an assumption.

Developments during the 1950s analyzed the demand for transactions balances rigorously from the standpoint of an individual who minimizes the costs of financing transactions by holding money balances and bonds, defined as interest-paying non-monetary financial assets. This analysis showed that the transactions demand for money depends negatively upon the bond rate of interest and that its elasticity with respect to the real level of expenditures is

less than unity. The original analyses along these lines were presented by Baumol (1952) and Tobin (1956). The following presentation draws heavily upon Baumol's treatment of the subject.

Developments since the 1950s have extended and broadened the Baumol–Tobin transactions demand analysis, without rejecting it. The most significant extension of this analysis has been to the case where there is uncertainty in the timings of the receipts and payments. The demand for money under this type of uncertainty is usually labeled as the precautionary demand for money and is the subject of Chapter 6.

4.1 The basic inventory analysis of the transactions demand for money

This section presents Baumol's (1952) version of the inventory analysis of the transactions demand for money. This analysis considers the choice between two assets, "money" and "bonds," whose discriminating characteristic is that money serves as the medium for payments in the purchase of commodities whereas bonds do not; hence, commodities trade against money, not against bonds. There is no uncertainty in the model, so the yield on bonds is known with certainty. The real-world counterpart of such bonds is interest-paying savings deposits or such riskless short-term financial assets as Treasury bills. Longer-term bonds whose yield is uncertain are not really considered in Baumol's analysis. Baumol's other assumptions are:

1 Money holdings do not pay interest. Bond holdings do so at the nominal rate R. There are no own-service costs of holding money or bonds, but there are transfer costs from one to the other, as outlined later. Bonds can be savings deposits or other financial assets.
2 There is no uncertainty even in the timing or amount of the individual's receipts and expenditures.
3 The individual intends to finance an amount $\$Y$ of expenditures, which occur in a steady stream through the given period, and already possesses the funds to meet these expenditures. Since money is the medium of payments in the model, all payments are made in money.
4 The individual intends to cash bonds in lots of $\$W$ spaced evenly through the period. For every withdrawal, he incurs a "brokerage (bonds–money transfer) cost" that has two components: a fixed cost of $\$B_0$ and a variable cost of B_1 per dollar withdrawn. Examples of such brokerage costs are broker's commission, banking charges and own (or personal) costs in terms of time and convenience for withdrawals from bonds. The overall cost per withdrawal of $\$W$ is $\$(B_0 + B_1 W)$.

Since the individual starts with $\$Y$ and spends it in a continuous even stream over the period, his average holdings, over the period, of the funds held in bonds B and money M are only $Y/2$. Hence, $M + B = \frac{1}{2}Y$.[1] Further, since the individual withdraws W each time and spends it in a continuous steady stream, and draws out a similar amount the moment it is spent, his average transactions balances M are $\frac{1}{2}W$. These propositions are shown in Figures 4.1 and 4.2. In Figure 4.1, for expenditures over one period, the triangle 0Y1 represents the amount of income that has not been spent at the various points of time within the period and 1YA is the amount that has been spent. 0Y1 equals $\frac{1}{2}Y$ over the period and would be held

1 The remainder ($Y/2$) passes out of his hands into sellers' hands.

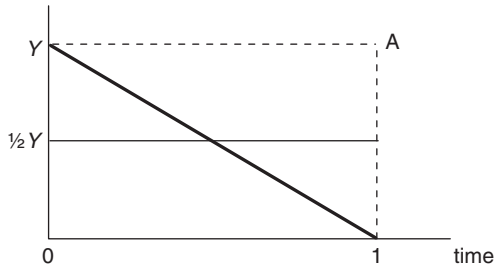


Figure 4.1

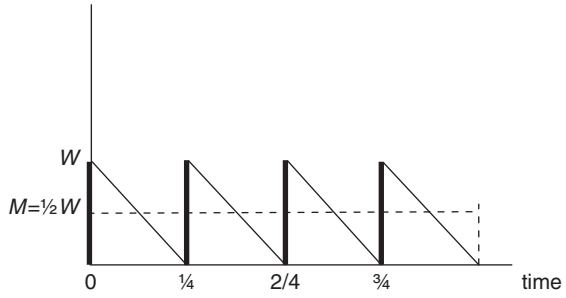


Figure 4.2

in either money or bonds. Figure 4.2 focuses on money holdings. To illustrate, assuming that the period is divided into 4 weeks, the amount $\$W$ is withdrawn at the beginning of each week and spent evenly through the week. The average money balances over the period are only $\frac{1}{2}W$, and, from Figures 4.1 and 4.2, the average bond holdings over the period are $(\frac{1}{2}Y - \frac{1}{2}W)$.[2]

Since the total expenditures of Y are withdrawn from bonds in lots of W, the number n of withdrawals is (Y/W). The cost of withdrawing Y from bonds is the cost per withdrawal times the number of withdrawals and is given by $[(B_0 + B_1 W)n]$. In addition, the interest foregone by holding money rather than bonds is RM. Since $M = \frac{1}{2}W$, this interest cost equals $RW/2$. The total opportunity cost C of financing Y of expenditures in this manner is the sum of the cost of withdrawing Y from investments and the interest foregone in holding average money balances of $(W/2)$. Hence,

$$\begin{aligned} C &= RM + (B_0 + B_1)n \\ &= RW/2 + B_0 \cdot Y/W + B_1 Y \end{aligned} \tag{1}$$

If the individual acts rationally in trying to meet his payments Y at minimum cost, he will minimize the cost C of holding transactions balances. To do so, set the derivative of (1) with

2 This amount, and our subsequent calculations of interest on the bonds held, implicitly assume that the withdrawals from bonds are continuous or almost continuous.

respect to W equal to zero. This yields:

$$\partial C/\partial W = R/2 - B_0 \cdot Y/W^2 = 0 \tag{2}$$

so that:

$$W = \left[2B_0 \cdot Y/R\right]^{½} \tag{3}$$

and

$$M^{\text{tr}} = ½W = (½B_0)^{½} Y^{-½} R^{-½} \tag{4}$$

where we have inserted the superscript tr to emphasize that (4) specifies only the transactions demand for money and does not include the money demand that would arise for speculative and other motives. (4) is called the *square root formula* in inventory analysis and has the easily identifiable form of a Cobb–Douglas function. In the present analysis, it specifies the demand for transactions balances for a cost-minimizing individual. The preceding demand function is clearly different from Keynes's demand function for transactions balances and, among other things, indicates that *the demand for transactions balances depends upon the nominal rate of interest*. The properties of this demand function, showing its response to changes in the real levels of expenditures, interest rates and prices, are discussed below.

Brokerage costs are the prices charged for brokerage services, which are commodities (i.e. "goods and services"), so that: let $B_0 = P.b_0$ and $B_1 = P.b_1$, where b_0 and b_1 are the elements of the brokerage charge *in real terms*, whereas B_0 and B_1 were nominal brokerage charges, and P is the price level. The reason for expressing brokerage costs in this way is that the brokerage services related to money withdrawals from earning assets are themselves commodities and, from a rigorous viewpoint, if the prices of *all* commodities double, the brokerage cost must also double. Hence, both B_0 and B_1 must be taken to increase in the same proportion as the commodity price level P.

Therefore, equation (4) can be rewritten as:

$$M^{\text{tr,d}} = (½b_0)^{½} P y^{½} R^{-½} \tag{5}$$

and

$$M^{\text{tr,d}}/P = m^{\text{tr}} = (½b_0)^{½} y^{½} R^{-½}$$

Therefore, the elasticities of the transactions demand for money with respect to y, R and P are:[3]

$$E_{m.y} = ½$$

$$E_{m.R} = -½$$

$$E_{M.P} = 1$$

$$E_{m.P} = 0$$

3 The elasticity of a variable y with respect to another variable x is defined as:

$$E_{y.x} = (x/y).(\text{d}y/\text{d}x)$$

In (5), since the elasticity of demand for real transactions balances with respect to *real* income is only ½, the demand for real transactions balances increases less than proportionately with the individual's real income due to economies of scale in the cost of money withdrawals from bonds. The elasticity of the transactions demand for money with respect to the nominal interest rate is $-½$: the higher the interest rate, the higher is the cost of holding funds in transactions balances and the lower is the demand for such balances. The elasticity ($E_{m.P}$) of the transactions demand for *real* balances with respect to an increase in all prices is zero, consistent with that derived for the general demand for money in Chapter 3. By implication, from (5), the elasticity $E_{M.P}$ of the transactions demand for nominal balances is 1.

Elasticity of the demand for nominal balances with respect to nominal income

We can now refine the implications of this analysis for the elasticity ($E_{M.Y}$) of the demand for *nominal* balances with respect to *nominal* income Py. Intuitively, since $Y = Py$, nominal income changes if either real income y or prices P change. Consequently, at rates of inflation close to zero, $E_{M.Y}$ will be approximated by $E_{m.y,}$ which is ½ in the above analysis. The higher the inflation rate, the more significant will be the influence of $E_{M.P}$, which is unity, so that in hyperinflation, $E_{M.Y}$ will approximate unity. Therefore, $E_{M.Y}$ will not be a constant over time but will vary between one-half and one, depending on real income growth relative to the inflation rate during the period under study. Both output and the price level change each period, so that, if their rates of change were roughly equal, Baumol's model implies that the estimated value of $E_{M.Y}$ should be near the mid-point of the range between 0.5 and 1.0. In fact, for developed economies with low rates of inflation, it is not unusual to find estimates of this elasticity somewhere near the middle (0.75) of the potential range. However, we should expect that economies with high (double-digit or higher) inflation rates would have higher estimated values of $E_{M.Y}$. In the limiting case of hyperinflation, the value of $E_{M.Y}$ should approach unity. These considerations imply that the estimated elasticity of demand for nominal transactions balances with respect to nominal income is likely to differ among sample periods if they have different growth rates of real output and prices.

4.2 Some special cases: the profitability of holding money and bonds for transactions

The above analysis incorporates the choice between holding money and the income-earning asset – "bonds" – to finance transactions. In exercising this choice, the individual will buy bonds only if he can make a profit from holding them; if he cannot, he will only hold money and equation (5) for the demand for real balances will not apply to him. For an analysis of this possibility, we need to derive the profit function from holding money and/or bonds.

As we have shown earlier through Figure 4.1, under Baumol's assumptions the individual spends his income Y in an even stream over the period and therefore holds $½Y$ on average in either money or bonds.[4] His average nominal holdings B of bonds are, therefore, equal to $(½Y - M)$, where, as before, M equals $½W$. The individual earns interest at the rate R on these bond holdings. The profit from holding either money or bonds equals this interest

4 The difference between these amounts (Y and $½Y$) represents the *average* amount ($½Y$) disbursed to other individuals in payments during the period. The total amount disbursed is, by assumption, Y.

income from holding bonds less the brokerage cost of money withdrawals from bonds.[5] That is, the profit π from using the combinations of money and bonds is given by:

$$\begin{aligned}\pi &= \text{interest income from bonds} - \text{brokerage expenses} \\ &= R \cdot B - (B_0 + B_1 W)n \\ &= R\{\tfrac{1}{2}Y - M\} - \{\tfrac{1}{2}B_0 Y/M + B_1 Y\} \end{aligned} \tag{6}$$

Maximizing (6) with respect to M yields the first-order maximizing condition as:

$$\partial\pi/\partial M = -R + \tfrac{1}{2}B_0 Y/M^2 = 0 \tag{7}$$

Hence, as in (4),

$$M^{\text{tr}} = (\tfrac{1}{2}B_0)^{1/2} Y^{1/2} R^{-1/2} \tag{4}$$

Further,

$$B^{\text{tr}} = \tfrac{1}{2}Y - (\tfrac{1}{2}B_0)^{1/2} Y^{1/2} R^{-1/2} \tag{8}$$

where the superscript tr on B emphasizes that this demand for bonds is only for transactions purposes. Hence, from (6),

$$\begin{aligned}\pi &= R\{\tfrac{1}{2}Y - (\tfrac{1}{2}B_0)^{1/2} Y^{1/2} R^{-1/2}\} - \{(\tfrac{1}{2}B_0)Y/[(\tfrac{1}{2}B_0)^{1/2} Y^{1/2} R^{-1/2}] + B_1 Y\} \\ &= \tfrac{1}{2}RY - (\tfrac{1}{2}B_0)^{1/2} Y^{1/2} R^{1/2} - \{(\tfrac{1}{2}B_0)^{1/2} Y^{1/2} R^{1/2}\} - B_1 Y \\ &= \tfrac{1}{2}RY - 2(\tfrac{1}{2}B_0)^{1/2} Y^{1/2} R^{1/2} - B_1 Y \end{aligned} \tag{9}$$

Simplifying, we get:

$$\begin{aligned}\pi &= \tfrac{1}{2}RY - 2RM - B_1 Y \\ &= (\tfrac{1}{2}R - B_1)Y \ - 2RM \end{aligned} \tag{10}$$

The last equation has an easy intuitive explanation: total interest income from holding money and bonds is reduced by the interest cost of holding money and the variable cost of withdrawing Y from bonds. Further, since the second term on the right-hand side is non-positive, the first term implies that, no matter what the level of income, it would not be profitable to hold bonds unless $R > 2B_1$.

In equation (10), π is non-positive if $R = 0$ or if the total brokerage charges exceed the income from holding bonds. The latter would occur if the brokerage costs are relatively high. Note in this regard that the brokerage costs include both the charges explicitly levied by financial institutions and any other costs of conversion from bonds to money. The latter include the time and inconvenience, etc. – sometimes referred to as the "*shoe-leather costs*" – of trips to the banks and other relevant financial institutions. These costs can be quite high in

5 Since the brokerage cost, as explained earlier, was $(B_0 Y/W + B_1 Y)$ and $M = \tfrac{1}{2}W$, the brokerage cost in terms of M is $(\tfrac{1}{2}B_0 Y/M + B_1 Y)$ in equation (6).

areas poorly served by financial institutions, as is common in the rural areas of developing economies and even sometimes of developed ones. They are dominant for individuals for whom the banks refuse to open accounts or for those who cannot meet the conditions – for example, acceptable references or minimum deposit balances – set by banks for opening or holding such accounts. In these cases, the individual will not find it profitable to hold bonds and will only hold money.

The profit from holding bonds in the transactions process is also non-positive if either income or/and interest rates are sufficiently low. Such considerations are relevant to relatively poor individuals or where the financial system and its regulation limits the interest rates that can be paid on bonds. In these cases, the individual's demand for bonds would again be zero.[6]

In cases of non-positive profits from holding bonds, i.e. $\pi \leq 0$, the demand for bonds would be zero and the optimal transactions demand for *nominal* balances would be:

$$M^{\text{tr}} = \tfrac{1}{2}Y \tag{11}$$

which has a unit income elasticity and a zero interest elasticity.

From (11), the transactions demand function for real balances m^{tr} is:

$$m^{\text{tr}} = \tfrac{1}{2}y \tag{12}$$

so that $E_{m.y} = 1$ and $E_{m.r} = 0$. In the nineteenth and early twentieth centuries the brokerage costs even for savings deposits were high,[7] while the incomes of most individuals were quite low, so that the money demand function was likely to be closer to (11) than to that implied by the inventory model. That is, the income elasticity of money demand was closer to unity, rather than to 0.5, and the interest elasticity was closer to zero, rather than to -0.5.

Even in the modern period, almost all economies have some individuals – usually those with lower incomes – with such a demand function. The more under-developed the financial system of the country or the local area, and the lower the incomes of the people, the more significant would be this factor. The inventory demand formula (5) thus tends to have limited validity for many less-developed economies and rural areas, and even for some segments of the population in the developed economies.

4.3 Demand for currency versus demand deposits

The above analysis does not really address the interesting question of the relative demands for currency, which is notes and coins, as against that for demand deposits. For this, we need to consider the cost, convenience and safety of holding and using currency as opposed to demand deposits in making payments, rather than the costs of conversion from "bonds" into these two forms of money. In the choice between using currency or demand deposits, demand deposits do have positive own costs of usage since they require some trips to the bank for making deposits and the banks often levy deposit and withdrawal charges on checks, whereas currency holdings do not involve any such charges for making payments from them.

6 Further, in perfect capital markets, the Fisher equation implies that the market interest rate is the sum of the real rate and the expected inflation rate. Therefore, in low-inflation environments, the market interest rate will be low.

7 Taking funds out of savings deposits or transferring them to a checking account required a trip to the bank branch, and bank branches were few and far between. Further, few individuals had bank accounts.

Further, the most common types of demand deposits do not pay interest. Consequently, currency involves lower own costs of usage, so that the optimizing individual will hold currency only and not demand deposits. This seems to be the case in many less-developed economies and especially in those rural areas poorly served by banks.

However, it is patently not the case in most developed economies or the urban sectors of developing economies that most individuals do not hold demand deposits, so there must be other considerations which are relevant to the choice between currency and demand deposits. The major one here for most individuals seems to be the relative safety of holding demand deposits as against that of holding currency.[8] The concern with theft and robbery if large sums were kept or carried in currency was a major reason for the origin and spread of deposit banking in eighteenth- and nineteenth-century Europe and continues to be a major determinant of the relative demand for demand deposits versus currency. The greater the concern with the safety and convenience of currency holdings, the lower will be the relative demand for currency balances. To illustrate, Japan, with an extremely low theft and robbery rate, is an economy in which ordinary persons do not normally hold demand deposit accounts but pay for most transactions in money. Conversely, persons carrying large sums in currency in the United States would be very concerned about their personal safety and the safety of these sums, and tend to prefer to hold demand deposits for meeting most of their transactions needs.

4.4 Impact of economies of scale and income distribution

Distribution of income

Consider the following two cases:

(A) An economy with n individuals, but with one having the whole of the national income Y and the rest with zero income. With zero income, the latter do not find it profitable to hold bonds and also do not hold money.
(B) An economy with n identical individuals, each having an identical income Y/n.

From (5), the nominal demands for transactions balances are:
For (A):

$$M_{\mathrm{A}}^{\mathrm{tr}} = (\tfrac{1}{2}B_0)^{1/2} Y^{1/2} R^{-1/2} \tag{13}$$

For (B):

$$\begin{aligned} M_{\mathrm{B}}^{\mathrm{tr}} &= n\left\{\tfrac{1}{2}B_0)^{1/2}(Y/n)^{1/2}R^{-1/2}\right\} \\ &= n^{1/2} M_{\mathrm{A}}^{\mathrm{tr}} \end{aligned} \tag{14}$$

Since $n \geq 2$, $M_{\mathrm{B}}^{\mathrm{tr}} > M_{\mathrm{A}}^{\mathrm{tr}}$. Hence, the equal distribution of incomes leads to a higher demand for real balances.

8 Another reason is the availability of very large denominations of notes. In the absence of these, it can be quite cumbersome to pay large sums in notes, implying high brokerage costs of using notes in such transactions. The central bank may not be willing to print notes of very large value to control illegal transactions and out of concern for the safety of the holders.

A more realistic scenario than either (A) or (B) would be one where a certain number of poor individuals have positive incomes but do not find it profitable to hold bonds. Their income elasticity of money demand would be one. In this case, the economy would have two types of individuals, ones with the usual Baumol money demand function and the others with $M = \frac{1}{2}Y$, so that the unequal distribution of incomes in this case produces greater money demand than under either (A) or (B). In the limiting case, imagine a scenario where incomes are equal but everyone is too poor to profitably hold bonds. This would imply the highest money demand, which would equal $\frac{1}{2}Y$.

Hence, provided all individuals have sufficient incomes to find it profitable to hold bonds, Baumol's model implies that the more unequal the distribution of incomes in the economy, the smaller will be the demand for real balances. However, this result may not hold if the unequal distribution of incomes leads to some individuals holding only money.

Economic development

The following analysis provides another example of the impact of economies of scale on the transactions holdings of money. Assume that, *ceteris paribus*, a fraction α of the population has enough income to hold transactions balances according to the inventory model, while the "poor" fraction $(1-\alpha)$ does not find it profitable to hold bonds for transactions purposes. The overall transactions demand for money per capita for the population is given by:

$$M^{\text{tr}} = \alpha(\tfrac{1}{2}B_0)^{\frac{1}{2}}Y_A^{\frac{1}{2}}R^{-\frac{1}{2}} + (1-\alpha)\cdot\tfrac{1}{2}Y_B \tag{15}$$

where Y_A is the income of each better-off person and Y_B is the income of each poor one. These money holdings have an income elasticity between ½ and 1. As the brokerage cost B_0 declines due to financial development or/and as the income of each poor individual rises sufficiently, α rises. This would raise the interest elasticity of overall transactions holdings in the economy and reduce their income elasticity. Therefore, economic and financial development should lead to a decrease in income elasticity and an increase in interest elasticity.

Further, note our earlier result that for a given interest rate, if brokerage costs are high and incomes low, as they are in many developing countries and especially outside the big cities, it would be unprofitable to hold bonds, so that the income elasticity of money demand would be one and the interest elasticity would be zero. In such a context, the average elasticity of money demand would be closer to one than to ½, which is implied by Baumol's model. This result would be reinforced in a context of inflation, since the price elasticity of nominal money demand is one.

4.5 Efficient funds management by firms

The preceding analysis was couched in terms of an individual but it can also be applied to firms. In the case of a firm with many branches, is it optimal for the firm to have centralized or decentralized money management? Centralized money management is here taken to mean a single account held by the firm as a whole, with the central financial department treating all the branches as one unit for its decisions on the amounts to be withdrawn each time. The amount withdrawn is then allocated among the branches. Decentralized money management means separate accounts and separate decisions on the amounts to be withdrawn at any one time.

Consider the case of a firm with total income or receipts equal to $\$Y$ and having n identical branches, each with income/receipts equal to $\$Y/n$. If it has centralized funds management,

with a single demand deposit account and investments from it into bonds, its cost-minimizing transactions balances would be as specified by (13). If it has decentralized funds management, with each branch holding its own demand deposit account and bonds, its transactions balances will be as specified by (14). The latter is larger the greater the number of branches.

Since centralized funds management implies lower transactions balances, it also implies higher profits. The efficient firm would, therefore, choose to centralize its fund management, all other things being the same. However, there are other factors that make at least partial decentralization of bank accounts desirable for firms. Among these are the convenience, bookkeeping and security aspects. Many firms consider these sufficiently significant to retain decentralized banking arrangements, with the balances being transmitted from the branches to a main account at periodical intervals or when they reach pre-specified levels. Hence, convenience and security reasons play important roles in the choice of the extent of centralization of deposits, as they do in the use of currency versus demand deposits.

In recent decades, the increasingly efficient electronic transfer and investment of funds have reduced brokerage costs and made it profitable for large firms to invest their surplus funds for periods as short as a day. Their desired end-of-the-day holdings of demand deposits may then be zero. Unpredictable withdrawals or deposits of funds can still occur, but these may be covered through overdraft facilities prearranged with the banks. In such a context, the actual holdings of demand deposits would be largely random. Such firms could still have positive currency demand but this would be largely in the nature of working or petty cash and depend upon considerations – for example, the unpredictable and uneven pattern of receipts and expenditures – other than those incorporated in the Baumol model.

4.6 The demand for money and the payment of interest on demand deposits

Many types of demand deposit accounts now pay interest. In order to properly consider these, assume that there are two assets, demand deposits and bonds, with each paying interest. Since currency does not pay interest, we exclude it from the definition of money in this section, so that money will equal demand deposits. The other assumptions of the model are as originally specified, including that the purchases of commodities can only be paid for by check drawn on a demand deposit account. As before, bonds are assumed to pay interest at the rate R, while demand deposits are now assumed to pay the rate R_D.

As in the preceding analysis, the average amount of demand deposits D is $W/2$ and that of bonds is $(\frac{1}{2}Y - D)$. The profit π from the use of money and bonds is:

$$\pi = R\left\{\tfrac{1}{2}Y - D\right\} + R_D D - \left\{\tfrac{1}{2}B_0 Y/D + B_1 Y\right\} \tag{16}$$

which yields the first-order maximizing condition as:

$$\partial\pi/\partial D = -R + R_D + \tfrac{1}{2}B_0 Y/D^2 = 0 \tag{17}$$

Hence,

$$D^{\text{tr}} = (\tfrac{1}{2}B_0)^{\frac{1}{2}} Y^{\frac{1}{2}} (R - R_D)^{-\frac{1}{2}} \tag{18}$$

$$B = \tfrac{1}{2}Y - (\tfrac{1}{2}B_0)^{\frac{1}{2}} Y^{\frac{1}{2}} (R - R_D)^{-\frac{1}{2}} \tag{19}$$

where:

$$E_{\mathrm{D}.(R-R\mathrm{D})} = -½ \tag{20}$$

$$E_{\mathrm{D}.R} = -½\left\{R/(R - R_{\mathrm{D}})\right\} \tag{21}$$

The demand for transactions balances now depends upon the interest rate *differential* $(R - R_{\mathrm{D}})$, and the elasticity of the transactions demand for demand deposits with respect to the differential in the interest rates is $-½$. However, this elasticity with respect to the bond rate of interest alone – that is, if the bond rate rises but the interest rate on demand deposits stays unchanged – is now $\left[-½\left\{R/(R - R_{\mathrm{D}})\right\}\right]$: the higher the interest rate on demand deposits, the higher is the elasticity of the demand for such balances. Since these elasticities are different, the impact on the demand for money of changes in bond yields will depend upon whether or not the interest rate on demand deposits is also changing.

4.7 Demand deposits versus savings deposits

As explained at the beginning of this chapter, non-checkable savings deposits can be viewed as a "bond" which pays interest but which cannot be directly used to make payments to others,[9] so that funds have to be transferred from savings accounts to checking accounts before a payment from them can be made by check. Prior to the advent of automatic banking machines and of telephonic and electronic transfers, a trip had to be made to a bank branch to transfer funds from savings accounts to a checking account or to obtain currency. Such a trip involved time and inconvenience, which are elements of the brokerage cost in the Baumol model. The proliferation of automatic banking machines and the general reduction in the banks' conditions and charges for such transfers have reduced this element of brokerage cost very considerably. The electronic transfer of funds among accounts handled through one's home computer has made this cost relatively insignificant.

Up to the 1960s, commercial banks also often imposed other costs, sometimes including a period of prior notice for withdrawal from savings accounts, for handling such transfers. The imposition of such notice has virtually disappeared. The result is that payments from savings deposits are now not very different in terms of costs and delays than from demand deposits.

For the following analysis, assume that savings deposits are the only kind of bond and the amount of savings deposits is designated as S. Since S replaces B in the analysis of section 4.2, the optimal ratio D/S in the context of that section is given by:

$$D/S = 1/\left\{½Y/D - 1\right\} \tag{22}$$

$$= 1/\left\{½(½B_0)^{-½}Y^{½}R^{-½} - 1\right\} \tag{23}$$

so that demand deposits fall with the decrease in brokerage costs. In the limit, $D/S \to 0$ as $B_0 \to 0$.

Historically, as the brokerage costs between demand deposits and savings accounts decreased, the proportion of balances held in demand deposits fell, so that this proportion is

9 There are now many types of savings accounts that pay interest and on which checks can be written. The difference between these and demand deposit accounts is not significant for our analysis since payments from both can be made by check. These can be treated as if they were demand deposit accounts.

currently less than 10 percent of M2 in the United States and Canada. Increasing familiarity in the handling of transfers between bank accounts from telephones and home computers is likely to further reduce this proportion.

The proliferation of automatic banking machines has also reduced the brokerage costs of transfers between currency and demand deposits, and also between currency and savings accounts. Therefore, as implied by Baumol's model, these banking facilities have allowed individuals to reduce their holdings of currency as against holding demand deposits and saving deposits. These banking facilities have therefore led to a decrease in both currency and demand deposits, so that the amounts held in M1 have fallen sharply.

4.8 Technical innovations and the demand for monetary assets

Recent decades have seen a considerable variety of innovations in the financial sector. The broad categories of these have been:

1 The creation of new types of financial assets and the increasing liquidity of some of the existing assets. These encompass institutional innovations such as interest-bearing demand deposits and checkable savings deposits, which did not become prevalent until the 1970s. They also include the issuance by banks of money market and other mutual funds, without a significant monetary brokerage charge for buying and selling such funds, and their divestiture into demand deposits at short notice. Such money market mutual funds, especially those sold by banks, became common only in the 1990s. Such innovations have shifted the transactions demand functions for currency, demand deposits and savings deposits.
2 Technical innovations in the deposit and withdrawal mechanisms and practices for various types of assets. These encompass the introduction of automatic teller machines (mainly in the 1980s) and telephonic and computer-based transfers of funds between accounts, beginning in the late 1990s but in common usage in this century. Debit cards are of this nature. They reduce the brokerage costs of using deposits, as against using checks, so that they reduce their transactions demand.
3 The development of "smart cards," which store nominal amounts, just as a coin or banknote does, and which allow the transfer of all or part of this amount to others at the point of the transaction without involving a third party such as a bank or a credit card company. Examples of these are certain types of telephone cards. Leaving aside the differences in technology and focusing on the economic nature, such cards are similar to coins and notes, which also embody value and allow the transfer of the whole or part of this value by the bearer to another person, the transaction proceeding with anonymity with respect to other parties. A rather insignificant difference is that paying with a larger note than necessary involves a reverse payment of "change," whereas the smart card allows transfer of the exact amount. The more important difference would be that a smart card with owner-authentication procedures built into it would prevent its theft to a much greater extent than is possible with currency, which can be used by the bearer without any authentication of proper ownership, so that the smart card would be more secure. This feature should make smart cards more attractive, and their use could replace that of both currency and checking accounts to a significant extent. In so far as both currency and smart cards constitute "value-carrying purses," the former being a non-electronic one and the latter an electronic one, it would be appropriate to lump them together in the total demand for "currency/purses" as against the demand for demand deposits, savings deposits, etc.

4 The development of digital cards, payments with which require the intervention of a third party such as a bank to verify, authorize and clear transactions over a network connection. These are more like checks or debit cards – whereas electronic purses are more like currency – and combine the advantages of checks with those of a credit or debit card. They leave a trail of transactions, which can be valuable for bookkeeping and security reasons.
5 The development of online payments, which allow payments to be made directly from a bank account to a payee. In the preceding analysis of the transactions demand for money, online payments reduce the monetary and non-monetary brokerage costs of using demand deposits and reduce their demand.

Hence, the very considerable – and continuing pace of – innovations in the financial industry in the past few decades have reduced the demand for currency, demand deposits and savings deposits, and have therefore shifted the demand functions for M1, M2 and the still wider definitions of money.

4.9 Estimating money demand

The inventory model of money demand implies that the alternative estimating log-linear forms of the transactions money demand equations are:

$$\ln M^{\mathrm{tr,d}} = \beta_0 + \beta_y \ln y + \beta_R \ln R + \beta_P \ln P \tag{24}$$

$$\ln m^{\mathrm{tr,d}} = \beta_0 + \beta_y \ln y + \beta_R \ln R \tag{25}$$

The model implies the elasticities $\beta_y = ½$, $\beta_R = -½$ and $\beta_P = 1$. But if the estimating equation had been formulated as:

$$\ln M^{\mathrm{tr,d}} = \alpha_0 + \alpha_y \ln Y + \alpha_R \ln R \tag{26}$$

the estimate of α_Y should lie between ½ and 1, with this value being larger the greater is the inflation rate relative to the real output growth rate. Further, in economies in which income, interest rate and brokerage costs are such as to make it unprofitable for most of the public to hold bonds for transactions purposes, the real income and nominal income elasticities would be closer to unity.

A rise in the interest rate causes two effects. One, it induces individuals to trade more often between money and financial assets, as the above inventory demand model shows. Two, for some individuals, who did not formerly trade between money and financial assets because of the unprofitability of doing so, the rise in the interest rate makes it profitable to undertake such trades, thereby increasing the interest elasticity of transactions money demand for the population as a whole. Hence, this elasticity should be non-linear.[10] Similar considerations applied to increases in income from very low levels would also cause the income elasticity to be non-linear.

In applying the above inventory analysis to the data collected on money balances, note that while the theory specifies average money balances held, the data is often collected as end-of-day (or other period) data. Further, the financially developed economies, with electronic

10 The empirical findings of Mulligan and Sala-i-Martin (2000) confirm this non-linearity of the interest elasticity and report that the interest elasticity of money is lower at low nominal interest rates.

transfers of funds at virtually zero brokerage costs, usually have one-day and overnight loan markets, in which firms using efficient cash management procedures can invest their excess money balances at the end of the day and others short of funds can borrow them. For *sweep accounts*, the banks themselves monitor the state of their customers' accounts at the end of each day and *sweep* the accounts of excess balances, investing them in overnight money market funds. In such a case, the customers need to ensure that they keep only the minimum desired balances; any amounts above or below these are lent or borrowed in the overnight or day-to-day loan markets or through loan arrangements such as overdrafts with their own bank (Bar-Ilan, 1990). Therefore, for customers with large balances and low transactions costs, the desired minimum transactions balances at the end of the day would be zero under the simpler versions of the inventory analysis. In other models, the amounts held by firms would be random, with a zero mean.

In a more realistic context, the customer, often an individual rather than a firm, would hold positive balances but these would be determined by institutional arrangements such as the minimum compensating balances banks sometimes require their customers to maintain in lieu of banking charges. Such considerations and the inventory model are more applicable to households' rather than to large firms' transactions demand for money.

Note that the data on money balances does not differentiate between those held for transactions and those held for speculative or other purposes. Hence, the preceding transactions demand elasticities provide only rough guides to the overall money demand elasticities. Further, financial innovation in recent decades is likely to have shifted the money demand function, so that the estimated elasticities would differ among different sample periods. Numerous empirical studies (see Chapter 9) confirm this finding.

Empirical findings

At a general level, the Baumol/Tobin analysis of transaction demand implies that the interest elasticity of money demand in developed economies with developed financial sectors will be negative. This has now been confirmed beyond any doubt by empirical studies on the overall demand for money (see Chapter 9).

The preceding analyses of transaction demand imply that the income elasticity of real balances with respect to real income is ½, their interest elasticity is $-½$ and the price elasticity of nominal money balances is one. Further, if it is unprofitable for the individual to hold bonds, because incomes and interest rates are relatively low and brokerage costs relatively high, the income elasticity of real balances with respect to real income is one, their interest elasticity is zero and the price elasticity of nominal money balances is one. Therefore, the decision to hold transactions balances involves two decisions: (i) whether to hold non-monetary interest-bearing financial assets; and (ii) how to allocate financial wealth between money and non-monetary interest-bearing financial assets. As income rises from low levels or brokerage costs fall with financial development, the average estimated income elasticity of real balances falls from a value close to one to a value close to ½. Empirical studies on the overall demand for money do usually estimate the income elasticity of money demand to be less than one (see Chapter 9).

For the usual income distributions with different incomes, the interest elasticity of transaction balances would be lower at low interest rates than at high interest rates, since more individuals in the population would find it profitable not to hold bonds, so that more of them would have zero interest elasticity of money demand. As interest rates rise, more

and more individuals would find it profitable to hold some bonds and substitute between money and bonds, so that the interest elasticity would increase towards ½. Therefore, money demand will be non-linear with respect to the interest rate, as will be the interest elasticity of the transactions demand. Mulligan and Sala-i-Martin (2000), using a cross-section sample of countries, confirm such non-linearity for money demand as a whole, with low interest elasticity at low income levels. For developing economies, as incomes rise, more and more individuals will find it profitable to use banking services and switch between (non-interest yielding) money and interest-bearing assets, so that the interest elasticity of money demand should rise and the income elasticity of nominal money balances should fall.

Since the transactions demand for money is only a component of the overall money demand, which cannot be separated in empirical estimation into its components, the empirical findings on money demand are left for detailed consideration to Chapter 9.

Conclusions

The basic conclusion of the inventory analysis for transactions demand is that, assuming positive profits from holding some bonds (including savings deposits) as part of the transactions portfolio, households will have economies of scale in holding demand deposits, and a negative interest elasticity with respect to the interest rate. This elasticity will differ depending upon whether or not interest is paid on demand deposits and upon the interest rate differential.

Innovations in electronic transfers and centralized control between the head office and branches, and between firms' branches and banks, have reduced the inconvenience connected with centralization and have thus promoted greater centralization of money management. Further, they have reduced brokerage costs for firms. In the limiting case where the brokerage costs per transaction at the margin tend to zero, the demand for demand deposits would tend to zero. As a consequence, the transactions balances held by firms relative to their revenues have fallen. Variations in these balances may be largely dominated by random factors in the case of large firms with efficient funds management in well-developed financial markets.

A consideration that leads to positive demand deposits being held are minimum compensating balances, often in lieu of transactions fees, sometimes required by banks. Such banking practices, as well as the number and sizes of branches, would be among the major determining factors determining the minimum holdings of money balances by individuals and firms.

The above discussion implies that the aggregate demand for transactions balances in the economy has three components. These are:

1. The demand by households and firms who do not find fund management with some investment in interest-bearing non-monetary assets ("bonds") profitable and hold only money. Such a component will exist in virtually any economy but may only be a significant part of the whole in economies with undeveloped banking and other financial facilities or in developing countries with low average incomes.
2. The transactions balances of those households and firms that find such financial management profitable. For these, the Baumol model would be applicable.
3. The demand by optimizing wealthy individuals and large firms for whom the variable part of the brokerage costs are almost zero. For these, the transactions balances are determined by factors not in the Baumol model. The relevant factors could be the requirement for payments in money to individuals in category 1 or for transactions for which the

requirement is to pay in currency (for example, for bus fares), or minimum balances required by banks to keep a demand deposit account.

The electronic, regulatory and institutional innovations in recent years have blurred the distinction between demand deposits and various near-monies, and thereby shifted the transactions demand for the former. The invention and use of devices such as electronic or smart cards is reducing the need to hold notes and coins for small expenditures, thereby reducing the demand for currency.

The inventory demand function is the core implication of the Baumol model. It was derived under rather special and restrictive assumptions. As this chapter has shown, relaxing these assumptions tends to change the implied elasticities of demand. However, in general, the qualitative conclusions remain: in the aggregate, the demand for real transactions balances increases less than proportionately with real expenditures, decreases with the yield on alternative assets, and does not change if all prices change proportionately.

Summary of critical conclusions

- The transactions demand for real balances has an elasticity of one-half with respect to real income and an elasticity of zero with respect to the price level.
- The transactions demand for nominal balances has an elasticity of one-half with respect to real income and unity with respect to the price level. Therefore, its elasticity with respect to nominal income is between one-half and unity.
- Efficient and centralized money management reduces the transactions demand for money.
- Financial innovations have reduced the demand for money, and have also made the transactions demand function unstable over time.

Review and discussion questions

1. Present Baumol's inventory analysis for the transactions demand for money.
2. Compare the cost minimization and the profit maximization approaches to the derivation of the transactions demand for money. What insights do we get for the transactions money demand from using the profit maximization approach that are not apparent from the cost minimization approach?
3. "Keynes's transactions demand function may not have been that unrealistic for the nineteenth and early twentieth centuries in Western, and other, economies." Discuss this statement.
4. "The nominal income elasticity of the transactions money demand is likely to differ among different sample periods for a given country and among countries." Discuss.
5. Explain why there is always a certain percentage of households that do not hold checking accounts even though they do use currency for transactions.
6. Derive the income and interest-rate elasticities in Baumol's inventory model of the transaction demand for demand deposits if interest is paid on demand deposits.
7. Why do modern societies use checking accounts when currency has lower brokerage costs than such accounts? Discuss.
8. Assuming that a firm has 25 branches, derive its demand for transactions balances and the income and interest rate elasticities if (a) each branch manages its funds separately, (b) there is central money management at the head office.

9. How would you incorporate security considerations/costs into the transactions demand model? What would this imply for the demand for currency in a relatively insecure urban environment (a) compared with a relatively safe one, (b) when owner-identified smart cards become available? Do these factors affect the demand for demand deposits? How would the proportion of currency to demand deposits be affected in these cases?
10. Can the transactions demand model be used to explain why financial innovations in recent decades have reduced the transactions demand for M1?
11. Are transactions demand models useless, as Sprenkle (1969) argued? If they are, how would you explain the demand for M1 or just for demand deposits in the economy?

References

Bar-Ilan, A. "Overdrafts and the demand for money." *American Economic Review*, 80, 1990, pp. 1201–16.

Baumol, W.J. "The transactions demand for cash: an inventory theoretic approach." *Quarterly Journal of Economics*, 66, 1952, pp. 545–56.

Mulligan, C., and Sala-i-Martin, X. "Extensive margins and the demand for money at low interest rates." *Journal of Political Economy*, 108, 2000, pp. 961–91.

Sprenkle, C.M. "The uselessness of transactions demand models." *Journal of Finance*, 24, 1969, pp. 835–47.

Tobin, J. "The interest-elasticity of transactions demand for cash." *Review of Economics and Statistics*, 28, 1956 pp. 241–7.

5 Portfolio selection and the speculative demand for money

This chapter presents the demand for money as a component of a portfolio in which the alternatives to holding money are bonds with uncertain rates of return. This topic initially arose from Keynes's contributions on the speculative demand for money but is now treated as part of portfolio selection analysis, with significant differences between these two approaches. Keynes's own approach was presented in Chapter 2 and has been superseded by the portfolio selection approach in this chapter.

Key concepts introduced in this chapter

- Portfolio selection
- Money as a riskless asset
- Normal probability distribution and its moments
- Expected utility hypothesis
- Von Neumann–Morgenstern utility function
- Risk aversion
- Efficient opportunity locus
- Constant absolute risk aversion
- Constant relative risk aversion
- Quadratic utility function

Keynes introduced the idea of speculative demand for money into the literature. This is a demand for money as an asset for holding wealth, rather than for transactions or precautionary purposes. In modern terminology, it would more appropriately be called the asset or portfolio demand for money. However, we shall continue to use the usual terminology and refer to it as the speculative demand for money.

The speculative demand for money arises because of the uncertainty of the yields on alternative assets. However, this demand is not the only part of money demand that is related to economic uncertainty. Another part is the precautionary demand for money, which is related to the uncertainty of incomes and consumption needs. The analysis of precautionary demand will be presented in the next chapter.

The assets considered in this chapter are money and bonds, with the term "bonds," as usual in monetary economics, referring to non-monetary financial assets and therefore encompassing the shares of corporations as well as other investments in a financial form. However, the analysis can be broadened to include physical assets as well. Physical assets are

generally not very relevant as an alternative to holding real balances in financially developed economies, but can be quite relevant in financially less-developed economies or for segments of the population which do not have easy access to non-monetary financial assets. Our analysis in this chapter will be on the choice between money and non-monetary financial assets.

Bonds are usually an uncertain vehicle for transferring purchasing power from the present to the future; both nominal and real yields on few, if any, assets are known in advance in a world beset, among other things, with the loss of purchasing power through inflation. This uncertainty of yields is not the only property or characteristic of financial assets. Such assets vary widely with respect to their acceptability in exchange, their maturity or marketability, their reversibility, their divisibility and the costs of their exchange into money.[1] Even in a world of uncertainty, the dominant determinants of the demand for financial assets by a small investor with very limited wealth may well be other than those related to the uncertainty of the yields on assets. Students themselves often fall into these categories, opting very often for a narrowly based portfolio of money balances and savings deposits, rather than opting for risky assets with their higher yields, because of the relatively high transactions costs for them of buying and selling bonds.

A significant factor in the choice among risky assets is the degree of lack of information about the factors determining their past and future yields and the costs of acquiring this information. These costs may be high, in terms of time, effort and money, relative to the increase in yields expected from better information. While the analysis presented in this chapter does not take account of the extent of the information available in forming expectations on asset returns, there is no reason to assume that the individual's choice among financial assets is not seriously affected by the extent of reliable information that is available on each asset and on the average of all assets.

However, the managers of large portfolios, whether of individuals, firms or financial institutions, do keep abreast of pertinent available information as a routine matter. For them, the problems of the indivisibility of assets are also less serious since the cost per unit of a financial asset will be relatively small in relation to the size of the portfolio. In large firms engaged in production or trade, and in financial institutions, the transfer into and out of a given asset and information-gathering are handled by the employees of the firm so that they are in the nature of fixed costs, while the variable transfer costs among assets tend to be relatively small. Therefore, the dominant considerations determining the short-run structure of large portfolios are the expected yields on the available assets and their perceived risks, rather than those imposed by indivisibilities, lack of information or significant variable transfer costs among assets.

The theories of portfolio selection explain the relationship between the yields or end-of-period values of assets and the investor's optimal portfolio. There are several types of theories of portfolio selection. The most common among these use portfolio selection analysis, especially its mean-variance version, which is based on the expected utility hypothesis (EUH). The analysis of this chapter is based on this hypothesis. Note that this analysis assumes that the consumption-saving decision has already been made and the question under consideration is of the optimal allocation of wealth among assets.

The individual may be concerned with the *yields* on his portfolio or with his undiscounted *terminal wealth* at the end of the investment period. Earlier treatments of this subject – for

1 Thus, even if all uncertainty with respect to the yields on assets were removed, a wide variety of assets with different characteristics and yields could continue to exist.

example, Tobin (1958, 1965) – generally followed the former.[2] However, the classical concept of assets as stores of value would emphasize their terminal (i.e. at the end of the investment period) net worth, which is the sum of the initial wealth and the yield on the assets. Further, writers concerned with explaining the general behavior of the individual in buying insurance or gambling (e.g. Friedman and Savage, 1948; Arrow, 1971) focus on the individual's terminal wealth, i.e. on the terminal net worth of assets. This is now the general pattern of analysis in monetary economics, and this chapter will follow this pattern.

Section 5.1 reviews the statistical relationships between the means and variances of the yields on the individual assets and those of the overall yield to the portfolio as a whole. Section 5.2 examines the individual's objectives for portfolio selection under uncertainty. Section 5.3 presents the concepts of risk aversion, indifference and preference. Section 5.4 presents the implications of expected utility maximization for attitudes to risk.

Section 5.5 derives the efficient opportunity locus relevant to portfolio selection. This locus corresponds to the budget line in the microeconomic theory of the consumer. Section 5.6 presents Tobin's famous analysis of the portfolio demand for money as an alternative to bonds. Section 5.7 extends our analysis to three common specific forms of the expected utility function: constant absolute risk aversion, constant relative risk aversion and quadratic.

While this chapter is devoted to the derivation of a portfolio demand for money, Sections 5.8 and 5.9 add somewhat heretical notes by asking the question whether there really does exist a stable demand function – or even a positive portfolio demand – for M1, and even M2, in the modern economy with well-developed financial markets. This is a legitimate question to ask since Keynes had proposed the existence of a speculative demand for money at a time when the financial system was less developed than the current one. In particular, the ordinary investor's access to easily available short-term bonds and money market mutual funds was almost non-existent. Since such instruments are now widely available at relatively low cost through banks and brokerage firms, the significance of a speculative demand component in M1, and possibly even in M2, in our financially well-developed economies is now doubtful, though broader definitions of money that include money market funds would still include it.

5.1 Probabilities, means and variances

Investors in financial assets possess information on the past performance of their assets and also have some pertinent knowledge of their current and likely future performance, the issuer of these assets, the performance of the economy, etc. The rational individual uses any such knowledge and intuition to form estimates of the likelihood of occurrence of each of these possible yields[3] so that the individual's subjective probability distribution of the yields on the available assets can be specified.[4]

2 An analysis based on the rate of return keeps constant the portfolio composition as the size of the portfolio changes. This may not be valid in all cases.

3 In experiments, the individual can always be made to refine his information so that the probability of occurrence that he attaches to any particular actual yield can be calculated as a unique number and not merely within a range. However, in the real world, this likelihood may be fudged and vague, or refined, depending upon the monetary stakes involved and the information at hand.

4 Note that in Keynes's analysis of the speculative demand for money, as set out above in Chapter 2, the individual had uni-valued expectations: he definitely expected a particular yield to occur so that his subjective probability attached to this yield was one. Further, this expected yield was taken to be independent of the actual one. That is,

Basing the individual's choices on the probability distribution would be analytically cumbersome unless the distribution could be represented by a small number of variables. For many distributions, the distribution can be described by the moments of the distribution: the expected yield, the standard deviation or variance, skewness, etc. For *normal* distributions, it is necessary to know only the expected return and standard deviation to describe the whole distribution. Hence, the individual whose choice is only among assets with the normal distributions of outcomes need not consider the probability of each of their outcomes separately but only their expected return and standard deviation. This is not necessarily true for other distributions and the individual may need knowledge of the other moments as well.

Our analysis will only consider the first two moments – that is, the expected value and the standard deviation – of the distribution of the expected end-of-period values of the assets and the portfolio. Analysis limited to these two moments is also known as mean-variance analysis. Such analysis implicitly assumes a normal distribution. For any asset i, designate the possible outcomes – that is, the anticipated values at the end-of-the-investment period – as v_{ki} and their associated probabilities as p_{ki}. The first two moments of the distribution (v_{ki}, p_{ki}) of the outcomes on asset i at the end of the investment period are calculated as:

$$\mu_i = \Sigma_k p_{ki} v_{ki} \tag{1}$$

$$\sigma_i^2 = \Sigma_k p_{ki}(v_{ki} - \mu_i)^2 \tag{2}$$

where:

p_{ki} = subjective probability of outcome k of asset i
v_{ki} = outcome k of asset i
μ_i = mathematical expectation of the outcomes of asset i
σ_i^2 = variance of the outcomes of asset i $(= \sigma_{ii})$
σ_i = standard deviation of the outcomes of asset i

The expected value (at the end of the investment period) of the *portfolio* and its standard deviation is calculated as follows:

$$\mu = \Sigma_i \mu_i x_i \tag{3}$$

$$\sigma^2 = \Sigma_i \Sigma_j \sigma_{ij} x_i x_j \tag{4}$$

$$= \Sigma_i \Sigma_j \rho_{ij} \sigma_i \sigma_j x_i x_j$$

where:

μ = expected value of (the outcomes to) the portfolio
σ = standard deviation of outcomes to the portfolio
ρ_{ij} = correlation coefficient of outcomes between the ith and jth assets
σ_{ij} = covariance of the outcomes to assets i and j
x_i = quantity of the ith asset
$\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$
$\sigma_{ii} = \sigma_i^2$
$\rho_{ii} = 1$

the individual was said to have "inelastic expectations" of his expected yield with respect to changes in the actual rate of interest in the market.

5.2 Wealth maximization versus expected utility maximization

Suppose the individual knows his subjective probability distribution of the outcomes on each asset and, by inference, on each possible combination of the assets he can hold in his portfolio. Two hypotheses on his objectives with respect to his portfolio are:

Hypothesis 1 (maximization of expected wealth):

The individual maximizes the expected value EW of his terminal wealth[5] *W (or the utility U(EW) of expected terminal wealth).*[6]

The argument in favor of this rule is that it would maximize the net worth of the portfolio over an infinite number of experiments or periods under unchanging conditions. It would, therefore, seem reasonable to accept such a hypothesis for an individual holding the same assets for an indefinite number of periods. However, most investors seem to be concerned with the value of the assets at the end of the current period or after a relatively small number of periods. Expectations of the values of the assets also constantly keep changing, so that the main justification for this rule, based on an infinite number of plays of the identical prospect, becomes inapplicable.

There are other arguments and counter-examples to the application of this rule, the most famous among these being the counter-example of the *St Petersburg Paradox*.[7] Historically, this paradox was used to discount the validity of the expected wealth hypothesis for the case where the deviation around the mean is important. A somewhat stronger[8] and more appealing argument against it can be formulated as an aspect of economic rationality or of the modern rational expectations hypothesis. This hypothesis is that the individual considers all the available information at his disposal where such information affects the future amount of his wealth. Therefore, if the probability distribution of future wealth has a non-zero standard deviation, the rational individual would consider it in his decisions and not follow the maximization of expected wealth (hypothesis 1 above), since this would mean ignoring

5 "Terminal wealth" refers to the wealth attained at the end of the period for which the individual expects to hold the assets.

6 Maximizing EW is the same as maximizing $U(EW)$ since EW is an admissible monotonic transformation of $U(EW)$.

7 The St Petersburg Paradox refers to the following gamble. Suppose an individual A tosses a coin until it comes up heads. If heads come up on the nth toss, he pays to B an amount y equal to (2^{n-1}) dollars. The expected value (Ey) of the payoff y to B is infinitely great since:

$$Ey = \{(1/2).2^0 + (1/2)^2.2^1 + (1/2)^3.2^2 + (1/2)^n.2^{n-1} + \cdots\} \to \infty$$

where, for each term on the right-hand side, the first component (the expression inside the parentheses) is the probability of getting an outcome, which is specified by the second component (the expression outside the parentheses).

Individuals are rarely, if ever, willing to offer very high prices for the opportunity to play as B in this game, thus refuting the expected wealth hypothesis. This refutation led to the origin of the expected utility hypothesis in the mid-eighteenth century and is used to buttress its usage even today.

8 While the St Petersburg Paradox was historically used to establish – and is still commonly used to justify – the expected utility hypothesis (EUH), note that the amounts usually offered to play B in experiments also tend to be far below the amounts which would be implied by the commonly used utility functions (such as the log of wealth) in the (EUH), thereby also creating doubts about the empirical validity of these functions or of the EUH itself.

information on the standard deviation.[9] Since a risk-averse individual dislikes risk, for which the standard deviation is the statistical proxy, the value that he would place on a prospect would be less than its expected value.[10]

Note that the maximization of expected wealth (i.e. of EW) is, for analytical purposes, identical to the maximization of the utility of expected wealth (i.e. of $U(EW)$) since EW is an admissible monotonic transformation[11] of $U(EW)$. Hence, the invalidity of the former also applies equally to the latter. Given the various arguments against this hypothesis (i.e. maximization of EW or $U(EW)$), it is rarely, if ever, used in portfolio selection theory and we will not pursue it further.

Hypothesis 2 (the expected utility hypothesis (EUH))[12]:

The individual maximizes the expected utility of his terminal wealth.

This hypothesis is now based on the *von Neumann–Morgenstern* (N–M) *set of axioms* and is known as the *expected utility hypothesis.* For an individual satisfying these axioms, a "*cardinal*"[13] or N–M utility function can be constructed. Further, the axioms imply that the individual will maximize the expected value, $EU(W)$, of the N–M utility function $U(W)$, rather than maximize $U(W)$ or $U(EW)$. The expected utility axioms and theorem can be found in many advanced microeconomics textbooks. For those interested in them, one of their versions is presented in Appendix 1 to this chapter.

Under the expected utility hypothesis, the individual will make his choices among assets by maximizing the expected utility of terminal wealth, i.e. maximizing $E(U(W))$, which is based on the probability distribution of his terminal (i.e. end of the investment period) wealth W, rather than by maximizing the utility of expected wealth $U(EW)$, which is based on only its first moment. This hypothesis makes the very plausible assumption that the individual "likes" expected wealth, meaning by this that, *ceteris paribus*, he prefers more to less of it. Portfolio selection analysis represents the riskiness of the portfolio by its standard deviation

9 The same arguments apply against a theory that would a priori bar the individual from considering the third and fourth moments of the subjective distribution of yields in his decisions when the probability distribution is not normal.

10 Further, the application of the rational expectations hypothesis requires that the individual also takes into account any other information in making his choice. This would include such factors as the limits on his wealth, the repetitiousness of the prospect, the chance of bankruptcy, etc.

11 A monotonic transformation $F(U(EW))$ of $U(EW)$ only requires that $F'(EW)$ have the same (positive) sign as $U'(W)$, so that $\partial F/\partial U$ has to be positive.

12 This hypothesis was first proposed by Daniel Bernoulli in the seventeenth century. The axioms for its modern version were first presented by von Neumann and Morgenstern in *The Theory of Games* (1946).

13 Economics has two major notions of the cardinality of the utility function. One of these – often called the Marshallian or Jevonian one – was proposed by the proponents (among whom were Menger, Jevons and Marshall) of cardinal utility analysis in the late nineteenth century. It required the constancy of the marginal utility of income. The validity of such an assumption is doubtful and the notion of Marshallian cardinal utility had been discarded in utility analysis by the 1930s. The other notion of cardinal utility is the von Neumann-Morgenstern (N–M) one, which is based on a set of axioms and does not assume the constancy of the marginal utility of income. In intuitive terms, they imply the constancy of the marginal utility of the probability of a positive outcome, which provides the yardstick for rendering utility cardinal, in a prospect in which there are only two outcomes, the positive one and zero.

(or its variance) and assumes that the individual dislikes risk, that is, he prefers less to more of it, *ceteris paribus*. Such an individual is said to be a *risk averter* and to possess *risk aversion*.

5.3 Risk preference, indifference and aversion

The theories of portfolio selection generally measure the *risk*[14] of holding assets by the standard deviation (or variance) of their outcomes, or of some function of them. Similarly, the riskiness of a portfolio is measured by the standard deviation σ (or variance σ^2) of wealth.

An individual's attitude to risk can be categorized into:

(i) Risk aversion

Portfolio selection theories assume that the individual is a risk averter if he likes expected wealth, so that $\partial U(EW)/\partial EW > 0$, and dislikes risk, so that $\partial U/\partial \sigma < 0$. These imply that the individual wants more than the expected value of a risky prospect before he would be willing to purchase it. Conversely, if the individual already owns a risky prospect, he would be willing to sell it for less than its expected value. That is, he would be willing to pay a premium to transfer the risk, e.g. of becoming ill or dying, to someone else, such as an insurance company.

A risk averter can have an increasing, decreasing or constant degree of risk aversion. These terms will be discussed later in this chapter.

(ii) Risk preference

An individual is a risk preferrer if he is willing to accept less than the expected value of a risky prospect to buy it and, if he already owns it, wants more than the expected value to be persuaded to sell it. Such an individual likes increases in expected return and in risk. The purchase of lottery tickets in the market place, if it is based only on an evaluation of the expected return and risk but does not include the joy and excitement of gambling, can be analyzed as a case of risk preference since the expected return is usually less than the cost of the lottery ticket.

(iii) Risk indifference

An individual is risk indifferent if he wants exactly the expected value of the prospect to be persuaded to buy it or sell it.

While risk preference and risk indifference lead to interesting scenarios, the plausible assumption for economic choices is that of risk aversion. This is the general assumption of the theories of portfolio selection, and is the assumption used in this chapter for analyzing the speculative demand for money.

14 Economic analysis of the first half of the twentieth century distinguished between *risk* and *uncertainty*. The context was one of risk when there existed objective probabilities on which the individual based his decisions. The context was one of uncertainty when there did not exist objective probabilities and especially when the information was vague and inadequate, so that decisions had to be based on subjective probabilities. Moreover, the individual took account of the inadequacy of the information on which they were based. This is the context to which virtually all economic decisions belong. However, the expected utility hypothesis ignores such elements and treats the subjective probabilities as if they were objective, so that it treats uncertainty as if it were risk.

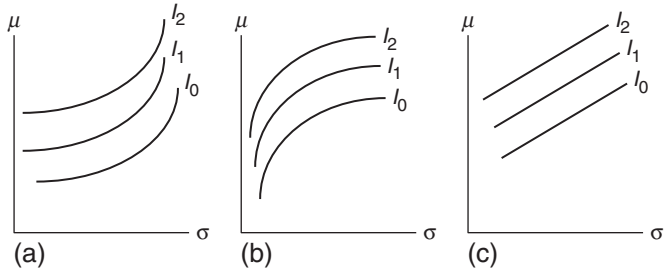


Figure 5.1

5.3.1 Indifference loci for a risk averter

Assume that a risk averter has chosen a portfolio with a particular combination of expected net worth and standard deviation of net worth. Suppose now that he is offered, *ceteris paribus,* an increase in the standard deviation of his net worth. Since he dislikes risk, he finds himself worse off by such an increase in risk. Therefore, if he is to remain indifferent between his initially chosen combination and one involving a higher risk, the expected value of his net worth must be simultaneously increased. Hence, indifference curves – the loci of (μ, σ) points between which he is indifferent – must be upward sloping, as in Figures 5.1 a, b and c, showing an increase in both expected net worth and risk along every indifference curve. The indifference curves shown in Figure 5.1a show increasing risk aversion as risk increases, so that $\partial^2\mu/\partial\sigma^2 > 0$; those in Figure 5.1b show decreasing risk aversion as risk increases, so that $\partial^2\mu/\partial\sigma^2 < 0$; and those in Figure 5.1c show constant risk aversion, so that $\partial^2\mu/\partial\sigma^2 = 0$.[15] The assumption of increasing risk aversion, as the risk to the portfolio increases, seems to be the most realistic one for portfolio allocation and is used further in this chapter.

The risk-averting individual prefers a higher level of expected net worth to a lower level, at any given risk. Hence, he prefers to be on a higher rather than a lower indifference curve, so that, for example, being on I_1 is preferred to being on I_0.

5.4 The expected utility hypothesis of portfolio selection

The expected utility hypothesis and the response to risk

Designate the individual's N–M utility function over terminal wealth W as $U(W)$. Assuming that the marginal utility of wealth $U'(W)$ is positive but decreasing at all levels of wealth, $U'(W) > 0$ and $U''(W) < 0$, where $U'(W) = \partial U/\partial W$ and $U''(W) = \partial^2 U/\partial W^2$. Such a utility function is shown by the curve marked $U(W)$ in Figure 5.2 and is concave to the origin. Wealth W is measured on the horizontal axis and the N–M utility of wealth is measured on the vertical axis. The individual has an initial wealth of W_0.

Suppose that this individual is offered an uncertain prospect L (where L stands for "lottery") which has two outcomes, W_1 and W_2, each equidistant (i.e. $|W_2 - W_0| = |W_0 - W_1|$) from W_0 and each with a probability of ½. If he accepts the prospect in exchange for W_0, he may

15 Figure 1c shows a constant positive degree of risk aversion as risk increases. If there were risk indifference, which translates to a constant zero degree of risk aversion, the indifference curves would be horizontal.

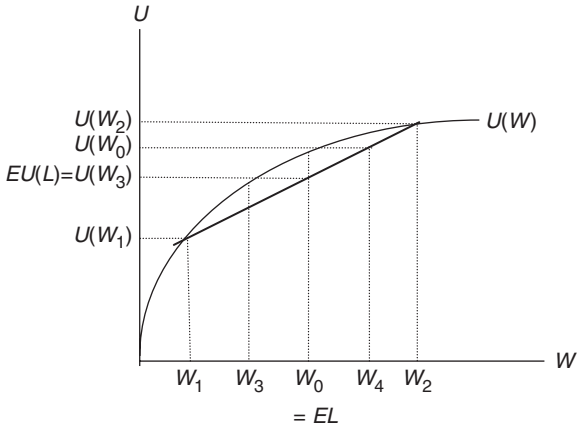


Figure 5.2

win ($W_2 - W_0$) or lose ($W_0 - W_1$), which are equal in absolute terms. The expected value of the prospect is EL, where:

$$EL = [½W_1 + ½W_2]$$

and its expected utility $EU(L)$ is:

$$EU(L) = [½U(W_1) + ½U(W_2)]$$

In Figure 5.2, the utility of having W_0 for sure is $U(W_0)$ while that of the prospect with the same expected value of wealth is $EU(L).U(W_0) > EU(L)$, so the individual will prefer the certain wealth W_0 to the risky prospect L, even though EL equals W_0. Hence, this individual is a risk averter. This result will hold if the individual has decreasing marginal utility over the relevant range of the prospect's outcomes.

Figure 5.2 also shows the maximum amount the individual will be willing to pay to buy the prospect. His initial utility is $U(W_0)$ for his initially certain wealth W_0. To persuade him to just buy a prospect with this wealth, the prospect must have an expected value EL such that $EU(L) = U(W_0)$. In Figure 5.2, this will be the prospect with $E(L) = W_4$.[16] Hence, $(W_4 - W_0)$ can be designated as the *risk premium* the individual – who does not yet own the prospect – will want to be paid to just accept the risk of the prospect. Alternatively, if the individual already owns a prospect having $EL = W_0$, he will be willing to sell it for the minimum sure amount of W_3 since $EU(L) = U(W_3)$ – thereby paying an *insurance premium* of $(W_0 - W_3)$ to get rid of the risk associated with his prospect.[17]

Therefore, the individual with decreasing marginal utility over a given range of wealth will be a risk averter for uncertain prospects involving outcomes within that range. It can similarly

16 This can be arranged by choosing appropriate probability p_1 of winning W_1 versus the probability $(1 - p_1)$ of winning W_2 such that $EL = p_1W_1 + (1 - p_1)W_2 = W_4$.

17 Therefore, a risk averter will be willing to buy insurance against the risks in his life and business – e.g. against death, disability, job loss, etc. – and be willing to pay a maximum amount of premium implied by

be shown that the individual with increasing marginal utility over a given range of wealth will be a risk preferrer for uncertain prospects involving outcomes within that range. Further, the individual with constant marginal utility over a given range of wealth will be risk-indifferent towards uncertain prospects involving outcomes within that range. These considerations led Friedman and Savage (1948) to argue that the individual who buys both insurance for some risks involving relatively small outcomes and lotteries with very much larger outcomes having a very small probability must have a segment with decreasing marginal utility, followed by a segment with increasing marginal utility at higher levels of wealth. Subsequent contributions modified this to the assertion that the utility function should be defined over $(W - W^c)$, where W is the individual's terminal wealth and W^c is his existing or customary level of wealth.

For the following portfolio selection analysis of the demand for money, the assumptions for the individual's preferences are that: (a) $U(W)$ is a von Neumann–Morgenstern utility function, (b) $U'(W) > 0$, and (c) $U''(W) < 0$. As we have shown above, the last assumption ensures that the investor is a risk averter.

5.5 The efficient opportunity locus

5.5.1 Expected value and standard deviation of the portfolio

For simplification, assume that the probability distributions are normal and therefore only the expected net worth and its standard deviation are relevant to the individual's decision. Further, assume that the individual can hold combinations of only two assets, X_1 and X_2, in his portfolio. Then, the first two moments of the frequency distribution of his terminal wealth, which is the value of the portfolio at the end of the relevant period, are:

$$\mu = \mu_1 x_1 + \mu_2 x_2 \tag{3'}$$

$$\sigma^2 = \sigma_1^2 x_1^2 + 2\rho_{12}\sigma_1\sigma_2 x_1 x_2 + \sigma_2^2 x_2^2 \tag{4'}$$

where:

x_1 = amount of asset X_1
x_2 = amount of asset X_2
μ = expected value of terminal wealth
σ^2 = variance of terminal wealth

The budget constraint on the holdings of the two assets is:

$$x_1 + x_2 = W \tag{5}$$

where W is the individual's initial wealth and the prices of the two assets have been normalized at unity to avoid continual usage of the price symbols.

his utility function and the risk in question. If a firm is willing to insure him for the risk for less than this amount, he will buy the insurance; if the required premium is greater than this amount, he will not buy the insurance.

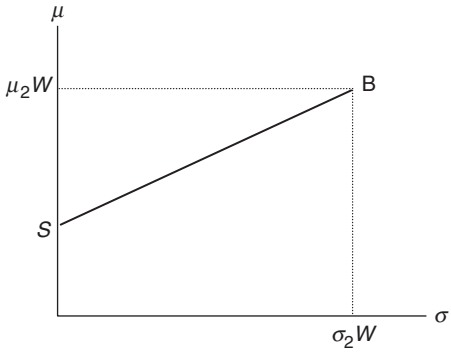


Figure 5.3

5.5.2 Opportunity locus for a riskless asset and a risky asset

Assume that the market offers a riskless asset S such that $\mu_s > 0$ and $\sigma_s = 0$, and a risky asset X_2 with $\mu_2 > \mu_s$ and $\sigma_2 > 0$. In this case, the opportunity locus for combinations with any risky asset would also be linear, as shown by the line SB in Figure 5.3. Intuitively, if the individual holds only the asset S, his yield will be $\mu_S W$, with $\sigma = 0$. If he holds only X_2, his yield to the portfolio would be $\mu_2 W$, with a risk of $\sigma_2 W$. If he holds half his wealth in S and half in X_2, he would have $\mu = (\frac{1}{2}\mu_s W + \frac{1}{2}\mu_2 W)$ and $\sigma = \frac{1}{2}\sigma_2 W$.

If the riskless asset M (say "money") also had a zero return – so that $\mu_M = 0$ – then the opportunity locus would be from the origin to point B.

5.5.3 Opportunity locus for risky assets

Appendix 2 presents the proper derivation of the opportunity locus between μ and σ for the various cases. Suppose the market offers only the two risky assets X_1 and X_2, with $\sigma_1, \sigma_2 > 0$. Define $x_1^* = x_1/W$ and $x_2^* = x_2/W = (1 - x_1^*)$, where x_1^* is the proportion of wealth held in the first asset. The following argument provides three cases for the shapes of the opportunity locus between μ and σ.

Perfect positive correlation, i.e. $\rho_{12} = 1$:

In this case,

$$\sigma^2 = \sigma_1^2 x_1^2 + 2\rho_{12}\sigma_1\sigma_2 x_1 x_2 + \sigma_2^2 x_2^2 \tag{6}$$

so that:

$$\sigma = \sigma_1 x_1 + \sigma_2 x_2$$

which, along with $\mu = \mu_1 x_1 + \mu_2 x_2$, gives a linear relationship between μ and σ. Such a relationship is shown by the line AB in Figure 5.4a. For the end-point A, representing the whole portfolio consisting only of the first asset, we have $x_1 = 1$ and $x_2 = (1 - x_1) = 0$, so

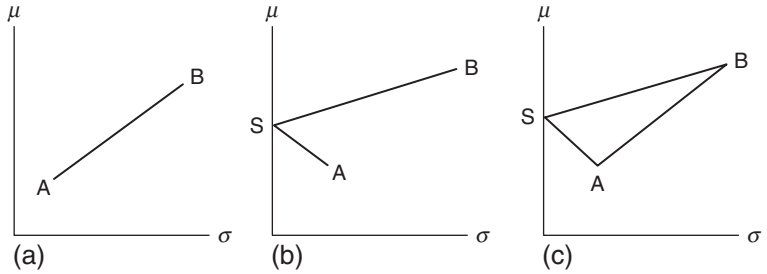


Figure 5.4

that $\mu = \mu_1$ and $\sigma = \sigma_1$. For the end-point B, representing the whole portfolio consisting only of the second asset, we have $x_1{}^* = 0$ and $x_2 = (1 - x_1) = 1$, so that:

$$\mu = \mu_2 = k_2 + k_1\sigma_2$$

which, along with $\sigma = \sigma_2$, provides the point B in Figure 5.4a.

Perfect negative correlation, i.e. $\rho_{12} = -1$:

In this case,

$$\sigma^2 = \sigma_1^2 x_1^2 - 2\sigma_1\sigma_2 x_1 x_2 + \sigma_2^2 x_2^2 \qquad (6')$$

so that:

$$\sigma = \sigma_1 x_1 - \sigma_2 x_2$$

Hence, $\sigma = 0$ for $x_1/x_2 = \sigma_2/\sigma_1$. Hence we can define a riskless composite asset X_3 for which $\sigma_3 = 0$. It would combine the assets X_1 and X_2 in the proportions given by $x_1/x_2 = \sigma_2/\sigma_1$, with $\mu_3 = \{(\mu_1\sigma_2 + \mu_2\sigma_1)W\}/(\sigma_1 + \sigma_2)$ and $\sigma_3 = 0$. In Figure 5.4b, a portfolio consisting only of X_3 is represented by the point S. Now, suppose that we have these three assets (X_1, X_2, X_3) in the portfolio. Combinations of only X_1 and X_3 in the portfolio yield $\sigma(X_1, X_3) = 0$ and $\mu(X_1, X_3) = \mu_1 X_1 + \mu_3 X_3$, with the linear opportunity locus AS in Figure 5.4b.[18] Similarly, combinations of only X_2 and X_3 in the portfolio yield $\sigma(X_2, X_3) = 0$ and $\mu(X_2, X_3) = \mu_2 x_2 + \mu_3 x_3$, with the linear opportunity locus SB in Figure 5.4b.

ASB represents the opportunity locus given by the combinations of the two risky assets with perfect negative correlation.

Opportunity locus for $-1 < \rho_{12} < +1$

In the common case where the two assets entail some risk and ρ_{12} lies between -1 and $+1$, the opportunity locus will be non-linear and will lie in the area enclosed by ASB in Figure 5.4c.

18 σ in $\sigma(X_1, X_3)$ and μ in $\mu(X_1, X_3)$ are being used as functional symbols so that their arguments can clearly indicate which assets are in the portfolio. However, σ and μ retain their definitions as being respectively the standard deviation and the expected return to the portfolio.

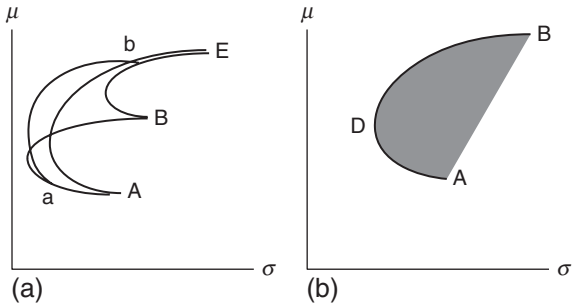


Figure 5.5

The closer ρ_{12} is to $+1$, the closer will this locus be to the line AB. When ρ_{12} differs from $+1$, there exist economies in risk from holding a diversified portfolio and the opportunity locus moves towards ASB.[19]

Opportunity locus for three risky assets

Now assume that three risky assets X_1, X_2, X_3 are available. Combinations of X_1 and X_2 alone, of X_1 and X_3 alone and of X_2 and X_3 alone, give the opportunity loci in Figure 5.5a as AB, AE and BE respectively. Further, consider combinations of the points *a* on AB and *b* on BE. These generate the locus *ab*. If we were similarly to take combinations of all the other points on AB, AE and BE, the opportunity locus would be the area enclosed by the curve ADB in Figure 5.5b. In the general case with more than three risky assets, the opportunity locus would have a similar shape.

Opportunity locus for one riskless and several risky assets

Consider, further, the possibility that the individual will wish to hold combinations of a riskless asset and the two risky assets. The individual can hold various combinations of the risky assets only or hold a combination of the riskless asset with some combination

19 This can be intuitively illustrated by the folklore that one should not put all one's eggs in one basket. It pays to divide one's eggs between two baskets as long as there is a possibility that when one of the baskets falls, the other one may not. If both baskets are guaranteed to fall simultaneously or not fall at all, e.g. when bound together, there is no advantage to separating the eggs between the baskets; the same number of eggs will break whether they are separated or not. In our formal language, the correlation coefficient of the baskets falling is one and there are no economies. However, if it is guaranteed or certain that when one basket falls, the other will not – as when the baskets are at the opposite ends of a pole carried across one's shoulders – separating the eggs into the two baskets will ensure that some eggs will with certainty escape breakage. In this case, the correlation coefficient of the baskets falling is (-1). In the intervening cases, the eggs should be divided between the baskets, the exact division depending upon the probability that when one basket falls, the other will not – that is, upon the correlation coefficient between the baskets falling. Similarly, the individual investor must consider the correlation coefficient of net worth between the assets since he may be able to reduce his risk through the diversification of the portfolio between the assets.

of the risky assets. Assuming that the return to the portfolio, when the total wealth is put in the riskless asset only, is given by OS, the efficient opportunity locus is shown by SGB in Figure 5.6. This locus has two segments, GB and SG. If the individual chooses a point on GB, he will hold only the risky assets and his demand for the riskless asset will be zero. If the individual chooses a point on SG, he will hold a portfolio consisting of the riskless asset and some combination of the risky ones, with only the riskless asset held at the point S. As he diversifies and holds increasing amounts of the risky assets, he moves towards G, with G representing a combination of only the risky assets. Note that the movement from S towards G represents an increase in the amount of the investment in the bundle of assets represented by G without a change in the relative composition of this bundle.

5.5.4 Efficient opportunity locus

The risk averter is concerned only with the part of the (μ, σ) combinations called the *efficient opportunity locus*, which gives him the highest possible μ for a given value of σ. Thus, in Figure 5.7, looking at the opportunity locus ADB for combinations of X_1 and X_2, with $-1 < \rho_{12} < 1$, the points on segment AD will be inefficient combinations of the two assets compared with those on segment DB; for example, point b on DB gives a higher μ at the given σ_0 than the combination a on AD. Therefore, AD can be disregarded for risk averters and the efficient opportunity locus for them is DB, which is non-convex to the origin.

5.5.5 Optimal choice

The preceding two subsections analyzed the risk averter's indifference curves and the opportunities open to him. Since such an individual prefers to be on a higher indifference curve to being on a lower one, he will prefer to be on the highest indifference curve that touches his opportunity locus. Such a curve will be that which is tangential to the efficient opportunity locus, and the individual's optimal combination of expected net worth and risk would be given by the point of tangency. Such points are shown by points a and b in Figures 5.8 and 5.9.

Demand among several riskless and risky assets and the speculative demand for money

The still more general case would have two riskless assets X_1 and X_2, with $\mu_1 > 0, \mu_2 > \mu_1$ and $\sigma_1 = \sigma_2 = 0$, and many risky assets. Since $\mu_2 > \mu_1$, and both assets have the same zero standard deviation, the rational individual would prefer to hold X_2 to X_1, so that his demand for asset X_1 with the lower return will be zero. To illustrate, if demand deposits and saving deposits in banks are both riskless but savings deposits pay a higher interest rate than demand deposits, as they normally do, the speculative demand for demand deposits – as well as currency – would be zero under this analysis. Similarly, if the money market mutual funds have higher expected return than savings deposits in banks, but can also be taken to be riskless because of their very short maturity, then the speculative demand for savings deposits would also be zero.

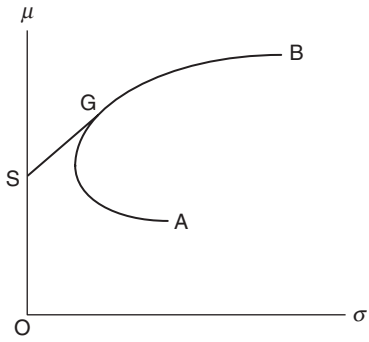


Figure 5.6

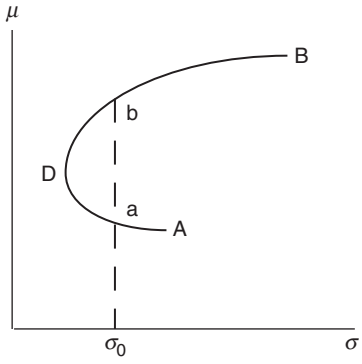


Figure 5.7

The optimal point *a* in Figure 5.8 is a combination of the riskless asset and the standard bundle of risky assets represented by point G.[20] The optimal point *b* in Figure 5.9 is a combination of the risky assets A and B only, while the riskless asset S is not held. A change in preferences that leaves the individual on the segment GB of the opportunity locus will change the relative demands for the risky assets A and B, but still without a positive demand for the riskless asset S. However, the individual whose initial optimal choice does not include the riskless asset may shift partly or wholly into the riskless asset, either because of an increase in his degree of risk aversion – which would shift the indifference curves – or because of an increase in the riskiness, or decrease in the expected net worth, of the risky assets – which would shift the efficient opportunity locus.

20 A shift in the degree of risk aversion that changes the slope of the indifference curves in the neighborhood of point *a* but leaves the optimal combination on the segment CG will change the demand for the riskless asset and the risky bundle. It will, however, leave the composition of the risky bundle unchanged. Such a shift in preferences was termed by Hicks a change in the degree of liquidity preference.

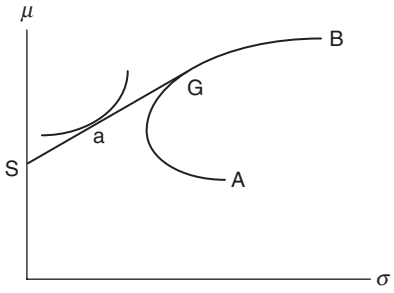


Figure 5.8

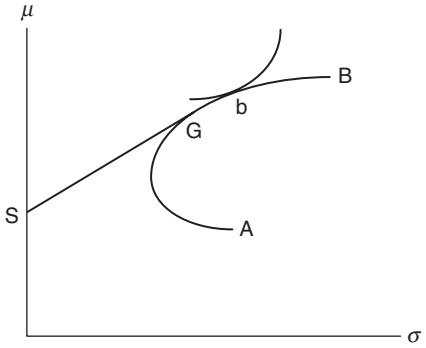


Figure 5.9

Volatility of the optimal portfolio

As discussed in Chapter 2, Keynes had argued that the individual's demand for money and bonds depends on subjective probabilities affected both by inadequate, and often conflicting, information, and by "herd behavior,"[21] so that their demand functions will be very volatile. In the context of the preceding portfolio selection analysis, periods of financial panic are likely to increase the individual's cautiousness (that is, his degree of risk aversion) while decreasing the expected return and increasing the perceived riskiness of many of the assets. The latter would shift the opportunity locus downwards, with a lower marginal return to risk. These periods of panic imply drastic decreases in the demand for risky assets and increases

21 Herd behavior is one where all or a majority of the participants in a market "run" (i.e. buy and sell in the same direction). It can be regarded as a case of "contagion" where the views of some carry along those of others on which way the stock market is expected to go. Stock markets often behave in this fashion, with their extreme manifestations being labeled, if positive, stock market "mania" or "exuberance" or, if negative, a "crash," "panic" or "nervous breakdown." Consequently, stock markets suffer from periodic overvaluations and undervaluations of stocks relative to their fundamental values. What happens during these episodes is sometimes attributed to "human nature" or "animal spirits," implying that it is beyond explanation by rational economic reasoning. The running of millions of wildebeest in Tanzania and Kenya in their migrations is sometimes used as an illustration of herd behavior.

in the demand for the riskless asset. Such shifts in demand may be self-reinforcing for some time since the fall in the demand for risky assets would lower their prices, causing a capital loss from the holdings of these assets. If these developments also cause the future expected net worth of the risky assets to fall and their expected riskiness to increase, the opportunity locus would shift down further, so that the optimal combination, with given preferences, will include still more of the riskless asset and even less of the risky assets. This process of decline in the net worth of risky assets and increased demand for the riskless asset could prove to be a *cumulative* one for some time.

In periods of optimism and a boom in asset prices, the opposite process is likely to occur. This would mean a decrease in demand for the riskless asset and an increase in demand for the risky assets, along with their prices. This movement could also prove to be a cumulative one for some time.

5.6 Tobin's analysis of the demand for a riskless asset versus a risky one

In an early form of the preceding analysis, Tobin (1958, pp. 65–86)[22] analyzed the demand for a riskless asset called "money" with a zero rate of return – that is, with a terminal (i.e. end-of-investment-period) value of unity and with zero standard deviation – as against a risky asset called *bonds,* with a positive return and a positive standard deviation. This analysis is presented in Figure 5.10, though with the difference that we will use savings deposits as our riskless asset – these deposits have a positive but still riskless rate of return.

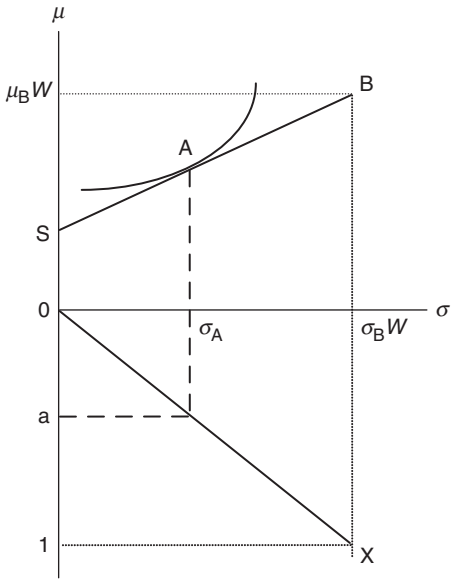


Figure 5.10

22 The part of Tobin's analysis that refers specifically to Keynes's arguments on the speculative demand was presented in Chapter 2.

Figure 5.10 has the expected value μ of terminal wealth on the vertical axis and the standard deviation σ of wealth on the horizontal axis. If the individual invests his wealth wholly in money, he will be at point S. If he invests only in bonds, he will at point B. The efficient opportunity locus is SB. In the bottom part of this figure, designate any point on the vertical axis as 1. The distance from 0 to any point on the segment 01 measures the proportion of wealth invested in bonds. The line from the origin down to the point given by $(1, \sigma_B W)$ is 0X.

If the individual chooses the point A, he will purchase σ_A, which implies, as shown in the bottom part of Figure 5.10, that he will place the proportion 0a in bonds and the proportion a1 in savings deposits.

We can use Figure 5.10 to investigate the effects of changes in the opportunity locus or the indifference curves. We consider three examples of the former.

(i) The first example considers a portfolio composed of a riskless asset with a positive return and a risky asset with a higher return. Assume that the tax authorities impose a lump-sum tax on a positive return to the portfolio, *without* a loss offset for investors incurring losses, in a manner such that the opportunity locus shifts down in a parallel fashion, as shown by the shift from SB to S′B′ in Figure 5.11.[23] Under this assumption, the particular shape of the indifference curves drawn in Figure 5.11 leads to an increase in the optimal purchase of risk to $0\sigma_{a'}$, so that the proportion invested in bonds increases to 0a′. This proportion is greater than the initial proportion 0a.

With S′B′ parallel to SB, the marginal rate of substitution $(\partial\mu/\partial\sigma)$ between the expected return and risk does not change. Hence, the substitution effect between them does not occur and the only effect in operation is the "income effect" – really a wealth effect in the context of portfolio selection – which could go either way, so that the after-tax optimal demand for

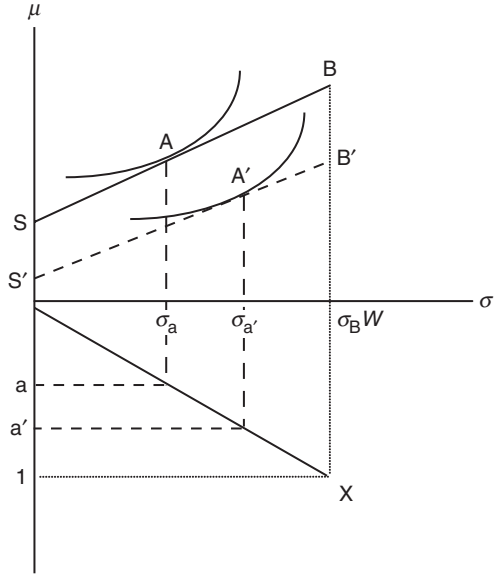


Figure 5.11

23 This assumes that the tax does not change the standard deviation of terminal wealth.

risk could be greater or smaller than the pre-tax proportion.[24] Figure 5.11 shows only one of these possibilities: at $\sigma_{a'}$, the net effect shown is an increase in the demand for risk. Since the riskiness of assets has not changed, the increased demand for risk translates into a higher demand for bonds. This is shown in the bottom half of Figure 5.11, which shows that the demand for bonds increases while that for the riskless asset falls. This example illustrates the behavior of an individual who strongly wishes to maintain the amount of terminal wealth and, with the decreased after-tax returns on his assets, has to purchase more bonds than before to do so. However, note that, depending on the tangency of the indifference curves to the new opportunity locus, the optimal point could have shifted to the left of point A and implied a smaller purchase of bonds.

(ii) The second example also deals with a portfolio composed of a riskless asset S with a positive return and a risky asset with a higher return. Assume that the tax authorities impose a 50 percent tax on a positive return and refund 50 percent of any negative return, so that the after-tax maximum amounts of μ and σ purchased through only holding bonds are cut down to those represented by point B′ in Figure 5.12. The after-tax return on the riskless asset is also cut down to half of the pre-tax return. The opportunity locus becomes S′B′, while the relevant line in the bottom part of Figure 5.12 becomes 0X′.

The analysis of this case is similar to that of the first example, except that, if the initial combination was to the left of B′, there is now greater likelihood of the optimal combination being at B′, which would mean a portfolio composed only of the risky asset. However, suppose that the optimal purchase of μ and σ is the same as before. This is the combination

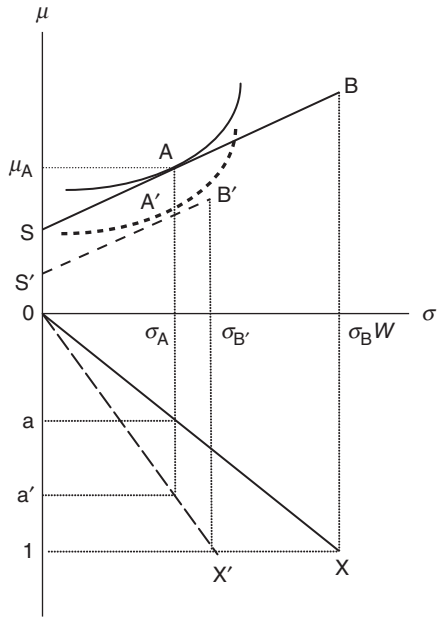


Figure 5.12

24 It would be smaller if the indifference curve tangential to S′B′ were such that the optimal point were to the left – rather than right – of point A.

indicated in Figure 5.12 by the point A′, which now implies a higher proportion (indicated by point a′) of the portfolio held in bonds and a smaller one in the riskless asset. Hence, although the tax reduced the after-tax return on bonds and the optimal combination of μ and σ remains the same, the amount invested in bonds increases because bonds now yield a lower return and risk. In this example, the substitution effect is again zero since the slope of the opportunity locus S′B′ is the same as that of SB, so that the income effect alone determines the changes in the desired holdings of money and bonds.

(iii) For the third example, assume, as in Tobin, that the riskless asset has a zero return. Tobin called such an asset "money." The relevant Figure is now Figure 5.13, with the initial opportunity locus being 0B. Further, assume that the tax authorities impose a 50 percent tax on the average rate of return on investments, without a loss offset for investors incurring losses, such that the post-tax opportunity locus becomes 0B′. Both income and substitution effects occur in this case. The optimal point in this figure becomes A′, so that the purchase of risk becomes $0\sigma_{a'}$, implying that the proportion invested in bonds falls to 0a′. This proportion is less than the initial proportion 0a, contrary to the situation shown in Figures 5.11 and 5.12. In this example, the nature of the indifference curves is drawn such as to imply an increased demand for the riskless asset.

The difference between Figures 5.11 and 5.13 is that the slope ($\partial\mu/\partial\sigma$) of the opportunity locus has shifted in the latter but not in the former. In Figure 5.13, the imposition of the tax reduces the marginal return to risk-taking – measured as $\partial\mu/\partial\sigma$ – so that the substitution effect comes into play and will cause a reduction in the optimal amount of the risk bought through bond purchases. Although the income effect could go either way, Figure 5.13 assumes that the two effects are in the same direction or that the substitution effect in favor of purchasing less risk dominates an opposing income effect. Since less risk is bought at A′, and the riskiness of bonds has not changed, the bottom part of this figure shows that the individual will invest a smaller part of the portfolio in bonds.

Figures 5.11 to 5.13 illustrate the use of the general utility function for deriving the demand for money, assumed to be riskless with zero or positive return, and risky assets. This analysis

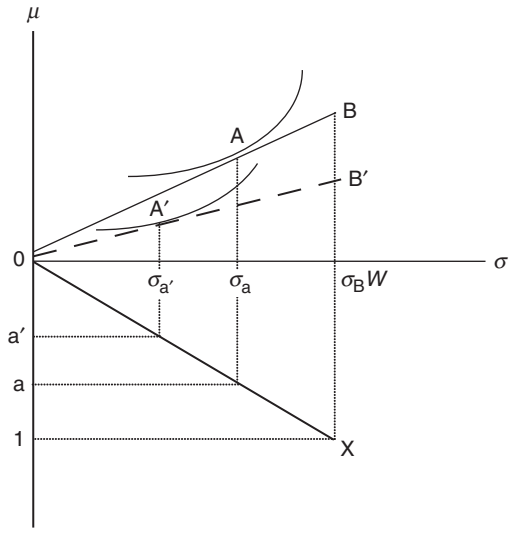


Figure 5.13

is diagrammatic and implies that the demand for money will depend on the expected return and standard deviation, as well as on the wealth to be allocated. Shifts in the opportunity locus because of tax changes or, more generally, because of changes in the perceptions of future risk and return – as occur periodically in the bond and stock markets – will alter the demand for money.

Since the preceding analysis has relied only upon indifference curves to represent preferences, its underlying utility function can be taken to be ordinal and need not be cardinal in the Neumann–Morgenstern sense. Current portfolio analysis is usually mathematical, based on the expected utility hypothesis, and tends to use specific cardinal utility functions. The next section presents this type of analysis.

5.7 Specific forms of the expected utility function

5.7.1 EUH and measures of risk aversion

If we look at the preceding indifference curve analysis, based on the mean–variance approach to portfolio selection, a suitable measure of risk aversion would seem to be $\partial\mu/\partial\sigma$. However, this measure does not directly relate to the form of the utility function that has wealth as its argument, and we sometimes want one that does so.

As shown above by Friedman–Savage analysis of Section 5.4 relating the slope of the utility function $U(W)$ to risk aversion, risk aversion is implied by the decrease in the marginal utility of wealth, so that one measure of risk aversion would be $U''(W)$, which is the change in the marginal utility of wealth. However, $U''(W)$ is not invariant to the admissible transformations of the utility function,[25] as we show in the next few paragraphs.

Cardinality of the von Neumann–Morgenstern utility function

As mentioned earlier, the *von Neumann–Morgenstern (N–M) utility function* is based on a set of axioms (see Appendix 1) which imply that, for an individual whose preferences obey the axioms, the utility function will be "*unique up to a linear transformation.*" That is, given an N–M utility function $U(W)$ for such an individual, we can construct another utility function $V(W)$ with identical indifference curves by a linear transformation, as in:

$$V(W) = a + bU(W) \qquad b > 0 \tag{7}$$

In constructing $V(W)$ from $U(W)$, a and b can be given any arbitrary values as long as b is positive. $V(W)$ is an equally valid utility function for an individual having $U(W)$, and the indifference curves for both $U(W)$ and $V(W)$ are identical. However, their derivatives do differ in value, since:

$$V'(W) = bU'(W) \tag{8}$$

$$V''(W) = bU''(W) \tag{9}$$

25 For ordinal utility functions, even the sign of $U''(W)$ can change under admissible (increasing) transformations of the utility function, so that it does not make sense to talk of decreasing marginal utility for them. For cardinal functions, the sign remains invariant under admissible (linear) transformations but the magnitude can change.

where $'$ designates the first-order derivative and $''$ designates the second-order one. Note that by the assumptions specifying a risk averter's preferences, U' and $V' > 0$ while U'' and $V'' < 0$. The above equations imply that:

$$V''(W)/V'(W) = U''(W)/U'(W) \tag{10}$$

Hence, U''/U' is not altered by admissible transformations of the N–M utility function. This property makes this ratio, as against U'' alone, appealing as a measure of the degree of risk aversion implied by the given utility function $U(W)$.

Measures of risk aversion

In (10), $U' > 0$, $U'' < 0$ and $U''/U' < 0$, so that the degree of risk aversion is usually measured by $[-U''/U']$, which gives a positive value. $[-U''/U']$ is called the *absolute degree of risk aversion*.

Since U' and U'' are unlikely to increase or decrease proportionately with wealth, $[-U''/U']$ will be affected by the amount of wealth. In order to ensure that the degree of risk aversion is independent of the level of wealth, another popular measure of the degree of risk aversion is $[W.U''/U']$. This measure is independent of the level of wealth, as well as of the arbitrary constants a and b.$[-W.U''/U']$ is called the *relative degree of risk aversion*.

The absolute and relative degrees of risk aversion can be calculated for any twice-differentiable utility function. While there is no a priori reason to expect that either of these will be constant for any given individual's particular utility function, the utility functions for which such constancy holds are analytically convenient to use and are therefore popular in economic analysis. We examine these in turn, and will follow them with a presentation of the quadratic utility function, which does not have the constancy of either the absolute or the relative degree of risk aversion but was used in some early expositions of the speculative demand for money.

5.7.2 Constant absolute risk aversion (CARA)

Constant absolute risk aversion (CARA) requires that:

$$-U''(W)/U'(W) = \gamma \qquad \gamma \geq 0 \tag{11}$$

where γ is the constant degree of ARA. Hence,

$$U''(W) = -\gamma U'(W) \tag{11'}$$

Since $U''(W)$ is a second-order derivative, integrating both sides of (11′) gives the utility function itself, though with two integrating constants which would have been dropped in the differentiation process. The utility function given by integrating (11) twice[26] is:

$$U(W) = a - b\exp(-\gamma W) \tag{12}$$

26 CARA requires that:

$$U''(W) = -\gamma U'(W)$$

Assuming that W is normally distributed with mean μ and standard deviation σ, as before, the expected value of this utility function is given by:

$$EU(W) = a - b[\exp(-\gamma\mu + \tfrac{1}{2}\gamma^2\sigma^2)] \tag{13}$$

Since $b > 0$ and $\gamma \geq 0$, maximizing the expected utility function (13) is equivalent to minimizing $[\exp(-\gamma\mu + \tfrac{1}{2}\gamma^2\sigma^2)]$ or maximizing:

$$(\mu - \tfrac{1}{2}\gamma\sigma^2) \tag{14}$$

In (14), substituting $\mu = \Sigma_i \mu_i x_i$ and $\sigma^2 = \Sigma_i \Sigma_j \rho_{ij}\sigma_i\sigma_j x_i x_j$ restates the decision problem as:

$$\text{Maximize} \quad \{\Sigma_i \mu_i x_i - \tfrac{1}{2}\gamma(\Sigma_i \Sigma_j \rho_{ij}\sigma_i\sigma_j x_i x_j)\} \qquad i, j = 1, \ldots, n$$

subject to:

$$\Sigma_i x_i = W \tag{15}$$

Note that x_i is the amount of the ith asset, not its proportion to wealth. Equation (15) represents one of the simplest decision frameworks in portfolio selection analysis, which is the main reason for its usage in such analysis. Note its main assumptions: the frequency distribution of terminal wealth W is normal, the individual's preferences satisfy the N–M axioms and the degree of absolute risk aversion is constant. The following analysis applies this analysis to the case of two risky assets.

CARA and the choice between two risky assets

Assume in the context of (15) that the individual's choice is between different quantities of the two risky assets X_1 and X_2 so that $n = 2$. His decision problem can be stated as:

$$\text{Maximize} \quad (\mu_1 x_1 + \mu_2 x_2) - \{\tfrac{1}{2}\gamma(\sigma_{11}x_1^2 + 2\sigma_{12}x_1x_2 + \sigma_{22}x_2^2)\}$$

subject to:

$$x_1 + x_2 = W \tag{16}$$

Hence, integrating both sides of this equation specifies that:

$$\ln U'(W) = k_1 - \gamma W$$

so that:

$$U'(W) = \exp(k_1 - \gamma W)$$

Integrating again,

$$U(W) = k_2 - (1/\gamma)\exp(k_1 - \gamma W) = a - b\exp(-\gamma W)$$

where k_1 and k_2 are constants of integration, $a = k_2$ and $b = (1/\gamma)\exp(k_1)$.

The Lagrangian function L for this problem is:

$$L = (\mu_1 x_1 + \mu_2 x_2) - \{\tfrac{1}{2}\gamma(\sigma_{11}x_1^2 + 2\sigma_{12}x_1x_2 + \sigma_{22}x_2^2)\} + \lambda[x_1 + x_2 - W] \tag{17}$$

where λ is the Lagrangian multiplier. The first-order conditions for maximizing L with respect to x_1, x_2 and λ are:

$$\partial L/\partial x_1 = \mu_1 - \gamma(\sigma_{11}x_1 + \sigma_{12}x_2) + \lambda = 0$$
$$\partial L/\partial x_2 = \mu_2 - \gamma(\sigma_{22}x_2 + \sigma_{12}x_1) + \lambda = 0$$
$$\partial L/\partial \lambda = x_1 + x_2 - W = 0$$

These conditions yield the optimal holdings of the two assets as:

$$x_1 = k_1 + k_2 W \tag{18}$$
$$x_2 = -k_1 + (1 - k_2)W \tag{19}$$

where:

$$k_1 = (\mu_1 - \mu_2)/\gamma(\sigma_{11} - 2\sigma_{12} + \sigma_{22})$$
$$k_2 = (\sigma_{22} - \sigma_{12})/(\sigma_{11} - 2\sigma_{12} + \sigma_{22})$$

so that the demand for each risky asset increases with wealth but the proportions of the two assets in the portfolio change as initial wealth W increases. These demand functions depend upon the expected returns and the variances and covariances.

CARA and the special case when a riskless asset is available

If X_1 were a *riskless* asset, i.e. with $\sigma_1 = 0$, while X_2 was still a risky asset, note that $\sigma_{11} = \sigma_{12} = 0$. Inserting these in the demand functions (18) and (19) implies that:

$$x_1 = W - (\mu_2 - \mu_1)/(\gamma\sigma_{22}) \tag{20}$$
$$x_2 = (\mu_2 - \mu_1)/(\gamma\sigma_{22}) \tag{21}$$

where x_2 (the demand for the risky asset) is independent of initial wealth, so that any increases in the initial wealth (beyond a certain level specified by $(\mu_2 - \mu_1)/(\gamma\sigma_{22})$) will be completely added to the riskless asset holdings. That is, for $W \geq (\mu_2 - \mu_1)/(\gamma\sigma_{22})$, $\partial x_1/\partial W = 1$ and $\partial x_2/\partial W = 0$. The amount of wealth held in the risky asset will stay unchanged as wealth increases beyond the amount specified by (21), so that, as the investor becomes wealthier, the proportion of his optimal portfolio allocated to the riskless asset will increase. This is far from the behavior pattern of most investors, who increase their holdings of the risky assets as their wealth increases. Financial markets now provide several riskless assets such as M1, savings deposits, money market mutual funds, etc., so that CARA implies that, beyond a certain amount of wealth, increases in wealth will be added to the riskless asset holdings, so that their proportion relative to wealth will increase. This prediction makes CARA especially unrealistic for portfolio allocation behavior. This result also holds if there is at least one riskless asset but several risky ones.

Hence, a CARA utility function, though analytically convenient, is not appropriate for the derivation of the speculative/portfolio demand function for M1, M2 or other

monetary aggregates, all of which include a riskless component, nor for the derivation of the portfolio demand for risky assets.

Limitation of CARA as a utility function for general portfolio selection

On a more general note, CARA implies that the investor becomes increasingly reluctant to take risks as he gets richer, so that he invests in increasing proportions of the *less* risky assets and decreasing proportions of the riskier assets. Intuitive knowledge of investor behavior indicates the opposite; most investors tend to be rather more cautious with limited wealth and become more willing to take chances as their wealth increases. Hence, CARA cannot be taken to be plausible for the choice among assets and its use in this sphere is mainly due to its analytical convenience. Constant relative risk aversion (CRRA) is more realistic in its implications for the demand for assets. We now turn to its analysis.

5.7.3 Constant relative risk aversion (CRRA)

Relative risk aversion (RRA) is specified by:

$$RRA = [-W.U''/U'] \tag{22}$$

Constant relative risk aversion (CRRA) is the requirement that:

$$[-W.U''/U'] = \beta \tag{23}$$

where β is the constant coefficient of RRA. Integrating both sides of this equation twice leads to the utility function (Cuthbertson, 1985, Ch. 3):

$$U(W) = a - bW^{1-\beta} \quad b > 0,\ \beta \neq 1 \tag{24}$$

$$= \ln W \qquad \text{for } \beta = 1 \tag{25}$$

where a and b are the constants of integration.[27] β is known as the Arrow–Pratt measure of relative risk aversion.

27 CRRA is often used in intertemporal consumption analysis, where the intertemporal utility function is assumed to be time separable and is specified as:

$$U(c_0, c_1, \ldots, c_T) = \Sigma_t u(c_t)/(1-\delta)^{-t} \qquad t = 0, 1, \ldots, T$$

where U is the intertemporal utility function and u is the period utility function. δ is the rate of time preference. u is assumed to be a CRRA function of the form:

$$u(c_t) = (c_t^{(1-b)})/(1-b) \qquad \text{for } b > 0, b \neq 1$$

$$= \ln c_t \qquad \text{for } b = 1$$

For these functions, the elasticity of substitution between consumption in any two periods is constant and equal to $1/b$.

The corresponding period utility function for CARA is:

$$u(c_t) = -(1/a)\exp(-ac_t) \qquad a > 0$$

In (25), $U(W) = \ln W$ is analytically very tractable and is probably the most popular form of the von Neumann–Morgenstern utility function. Its degrees of absolute and relative risk aversion are:

$$ARA = 1/W$$

$$RRA = 1$$

so that the absolute degree of risk aversion decreases as the investor gets richer. This is also true of the general form of the CRRA utility function.

CRRA and the choice among risky assets

If there are only risky assets, maximizing the expected value of the CRRA utility function subject to the budget constraint yields the demand function for the *i*th risky asset as:

$$x_i/W = k_i\mu \quad i = 1, 2, \ldots, n \tag{26}$$

where:

$$k_i = k_i(\beta, \boldsymbol{\mu}, \boldsymbol{\sigma}_{ij})$$

The variables in bold type indicate the vectors of the relevant variables (Cuthbertson, 1985, Ch. 3). Equation (26) implies that the proportions of the risky and riskless assets in the portfolio remain unchanged as wealth increases, so that the elasticity of the demand for each asset with respect to wealth is unity.

CRRA and the demand functions for monetary aggregates

If the first asset is riskless while the others are all risky, the demand function implied by (26) for the riskless asset is:

$$\begin{aligned} x_1/W &= 1 - \Sigma_i(x_i/W) \qquad i = 2, \ldots, n \\ &= 1 - \Sigma_i k_i \mu_i \end{aligned} \tag{27}$$[28]

The asset demand functions (26) and (27) are homogeneous of degree one in W, so that the proportions of the riskless and each of the risky assets in the portfolio remain constant as the portfolio grows. Alternatively stated, the individual's portfolio demand for each asset, whether riskless or risky, has an elasticity of unity with respect to his wealth, which is much more plausible than the implication of CARA for the demand for the monetary aggregates. However, casual intuition does suggest that in financially developed economies the individual's demands for currency and even demand deposits do not increase proportionately with his wealth, so that CRRA may not be suitable for deriving the portfolio demand for M1. This is not necessarily so for the other components, M2, M3 and so on, for which CRRA may perform better. The usefulness of CRRA for predicting the demand

28 See Cuthbertson (1985, Ch. 3) for the derivations of this and earlier results on the CRRA utility function.

functions for M1, M2 and wider monetary aggregates will, therefore, vary with the aggregate in question, but will not be appropriate for their common element, which is M1. However, as we discuss later in this chapter, the portfolio demand for M1 may, in fact, be zero in the modern economy.

5.7.4 Quadratic utility function

Besides CARA and CRRA, the third type of utility function in common usage in portfolio selection theory is:

$$U(W) = a + bW - cW^2 \qquad a, b, c > 0 \tag{28}$$

Hence,

$$EU(W) = a + b\mu - cE(W^2) \tag{29}$$

where, by the definition of σ^2,

$$\begin{aligned}\sigma^2 &= E(W-\mu)^2 = E(W^2 - 2W.\mu + \mu^2)\\ &= EW^2 - 2EW.\mu + \mu^2 = E(W^2) - 2\mu^2 + \mu^2\\ &= (EW^2) - \mu^2\end{aligned}$$

Therefore,

$$E(W^2) = \mu^2 + \sigma^2 \tag{30}$$

Hence, from (29) and (30),

$$EU(W) = a + b\mu - c\mu^2 - c\sigma^2 \qquad a, b, c > 0 \tag{31}$$

The individual maximizes (31) (subject to the budget constraint) with respect to μ and σ. To derive the asset demand functions, substitute the equations for μ and σ in (31) and maximize the resulting expected utility function with respect to the quantities of the assets. Since the quadratic utility function does not possess the CRRA property, the proportions of the assets will not remain constant as wealth increases.

The quadratic utility function is a second-degree polynomial in W. Such a polynomial transforms to an expected utility function in the first two moments of the probability distribution of wealth. Polynomials of higher degree can also be used as utility functions. These would bring into the expected utility function the other moments of the probability distribution, which is desirable, but they are usually quite intractable for analysis.

Limitations of the quadratic utility function

Since the utility function must have positive marginal utility of wealth, it must satisfy $\partial U/\partial W = b - 2cW > 0$, which requires that $W < 2c/b$. This restriction severely limits the range over which the quadratic utility function is applicable.

Further, for the quadratic utility function in (28), the absolute degree of risk aversion (ARA) is given by:

$$ARA = -U''(W)/U'(W) = 2c/(b - 2cW)$$

Since $(b - 2cW) > 0$, the ARA for the quadratic utility function is positive and increasing in W, so that as the investor gets wealthier, he becomes more risk averse. This increase – rather than a decrease – makes the quadratic utility function even less appealing than the CARA, but, as shown above, the latter has its own limitations.

To conclude, the CARA and quadratic utility functions are inappropriate for the analysis of the portfolio demand for money and bonds, while CRRA is more suitable.[29]

5.8 Volatility of the money demand function

Note that the speculative demand for money and the coefficients of its independent variables will depend upon the means, the standard deviations and the correlation coefficients of the expected terminal values of the assets, for all of which the subjectively expected future and not the past actual values are the relevant ones. Keynes (1936, Ch. 13) argued that the expected bond yields and equity prices depend on the mood of the market participants and their perceptions of the future, which are often based on very limited information and subject to the "herd instinct." These elements of the financial markets are as much in evidence today as in Keynes's day, as the daily movements of the stock market indices clearly indicate.

These arguments imply that the demand functions for bonds and equities are constantly shifting, so that they could not be properly estimated or, if estimated, would be worthless – unless the nature of the shift could be specified and adjustments made for it. If we follow Keynes's analysis, presented in Chapter 2 above, of bulls and bears for the speculative demand for money, this demand function must also be similarly volatile.

Empirical studies, reported in Chapter 9, do not generally estimate a speculative demand for money separately from the demand for real balances as a whole. While these studies sometimes show instability of the demand function for money as a whole over time, they do not show the sort of high volatility, due to sudden shifts in expectations, suggested by Keynes. Further, although considerable shifts in the estimated functions have been observed by studies using an annual or quarterly date in recent years, they seem to have been mainly due to innovations in transactions technology.

5.9 Is there a positive portfolio demand for money balances in the modern economy?

The modern economy with a well-developed financial sector has a plethora of financial assets that are as riskless as currency and demand deposits, or close enough in riskiness not to matter much to the individuals in the economy. Among these assets are various types of savings deposits, term deposits, certificates of deposit (CDs) and very short-term money market instruments. Since they pay higher returns than M1 without an accompanying higher

29 However, this assertion does not necessarily mean that its implications for asset demand functions are valid. It is commonly believed that the wealthier investors are, *ceteris paribus*, less risk averse. If so, their degree of relative risk aversion would not be constant but decreasing.

risk, M1 will not be part of the efficient opportunity locus and will not be demanded for speculative purposes (Chang *et al*., 1983). Similarly, there will be a portfolio demand only for the savings deposits component of M2, as long as other riskless assets in the economy do not dominate these deposits in expected return.

Consequently, in economies with a variety of riskless assets which are riskless in nominal terms but which do not directly circulate as a medium of exchange and are therefore not part of narrow money, the speculative demand for M1 (which is the medium of exchange but pays a lower rate of interest) would be non-existent or confined to those individuals who do not have access to other riskless assets at a low enough cost. Hence, the speculative demand model may be generally applicable – and the speculative demand for narrow money may be positive – in economies with poorly developed financial institutions and markets, where other riskless assets do not exist or do not dominate over money in their return. However, this model no longer seems applicable to the M1 holdings of the common households, firms and financial institutions in the modern developed economy. In such a context, while there may be a significant and large speculative demand for certain forms of savings deposits and for money market instruments, there need not be a significant demand for currency and demand deposits.

In terms of the general evolution of the financial sectors in Western economies, the increasing proliferation of banks since the 1950s, accompanied by a considerable increase in the ease of transfer from savings accounts to checking accounts, especially in the banks, have brought about an increasing dominance of the net return (over transfer costs) on savings deposits and a continuing increase in the proportion of savings to M1, with a corresponding increase in the M2/M1 ratio. The innovation of automatic tellers and their spread in the 1980s hastened this movement, so that M1 now tends to be quite small relative to M2 in economies with developed financial systems.

In more recent years, in the North American and European economies, it has become possible to buy and sell almost without notice, and without significant brokerage costs, various types of mutual funds through commercial banks. Among these, the money market funds, with investments in Treasury Bills of a month's or a few months' maturity, are virtually riskless and offer a higher yield than most savings deposits, which are often held in the same financial institution. The preceding analysis implies that the ratio of M2 (which excludes such money market funds) to a still wider definition of money is also likely to decline.

The inventory analysis of transactions demand for money, including M1, and bonds in the preceding chapter had implied positive demand levels for both these assets. This analysis had shown the brokerage costs of transactions in and out of bonds, as well as the nominal interest rate, to be a major determinant of these demands. This chapter provides the portfolio demand for these assets, but ignores brokerage costs and the medium-of-payments role of money. Few models combine both the transactions and portfolio demands in a single, coherent and tractable analysis. However, we can still intuitively examine the impact of the medium-of-payments role of money and of brokerage costs on the portfolio demands for money and bonds.

First, consider the medium-of-payments role of money in the context of portfolio switches. In this role, bonds trade against money, but not directly against commodities, and sales of one type of bond do not occur simultaneously with the purchase of another type of bond. The delay may be due to institutional practices[30] or to the investor's inertia in making purchases even after the funds have appeared in the investor's accounts. Money is held in the interval

30 For example, many brokerage firms credit customers' accounts with the funds some days after a sale of bonds occurs.

between sales and purchases, so that positive balances, related to transactions, are held in the management and switches of portfolios. In addition, switches among bonds involve two transactions, each incurring transactions costs, while a switch from money to bonds involves only one transaction, so that for very short holding periods of some bonds it might be profitable to hold money rather than bonds.

Therefore, overall, the portfolio demand for money would depend on the risk and return factors encompassed in the mean–variance analysis and the nature of the payments system relevant to switches among risky assets, as well as brokerage costs in such switches. In any case, in the presence of several interest-paying riskless assets, the portfolio demand for non-interest-paying money balances is likely to be relatively small and more likely to be significant for small investors than for financial institutions themselves.

Conclusions

Keynes introduced the speculative demand function for money into the literature, and Friedman embedded aspects of the demand for money as a temporary abode of purchasing power in neoclassical money demand analysis. This role of money occupied centre stage in the analysis of the demand for money for several decades and was used to show the dependence of the demand for money on the rates of interest. The varieties of the analytical developments discussed in this chapter are testimony to the importance of the speculative motive in the literature on monetary economics.

The demand function for speculative balances derived in this chapter includes, among its arguments, wealth (rather than current income) and the expected yields (rather than actual yields) on the available assets. Such a function would be stable in an unchanging environment but could be unstable in periods of volatility in the bond and stock markets, changes in financial regulations, innovations in the characteristics of existing assets and the creation of new ones, and changes in the payment mechanisms.

The applicability of this analysis to the demand for money in the modern economy needs to be carefully reconsidered. Keynes's analysis was based on only money – the only available liquid and riskless asset – and consols. In such a context, the uncertainty of the terminal value of consols would create a demand for money in the economy. But developed economies since Keynes's day have created a wide variety of riskless assets that pay positive rates of interest. In these economies, economic units with access to such assets would prefer to hold them as temporary abodes of purchasing power for speculative purposes rather than holding currency and demand deposits, which either do not pay interest or pay lower rates of interest than riskless savings accounts. Therefore, the demand for M1 – and even for wider definitions of money such as M2, whose savings deposit component now faces competition from money market funds – must come from transactions and precautionary motives rather than from speculative motives. The transactions demand analysis was presented in Chapter 4 and the precautionary demand analysis will be presented in Chapter 6. A related topic is that of the buffer stock role of money, which is analyzed in Chapter 6.

Appendix 1

Axioms and theorem of the expected utility hypothesis

This appendix gives one version of the N–M axioms and theorem (Handa, 1983). Define a prospect (also called a lottery or gamble) (x, p) as one having the outcomes $x(=x_1, x_2, \ldots, x_n)$

with respective probabilities of $p(= p_1, p_2, \ldots, p_n)$. The prospects (q, r) and (z, r) are defined correspondingly. $\preceq$ is the symbol for "not preferred to" and $=$ stands for "is indifferent to."

Axiom 1 (Ordering)

$(x,p) \preceq (y,q)$ *or* $(y,q) \preceq (x,p)$.

If $(x,p) \preceq (y,q)$ and $(y,q) \preceq (z,r)$, then $(x,p) \preceq (z,r)$.

Axiom 2 (Continuity)

If $(x,p) \preceq (y,q)$ and $(y,q) \preceq (z,r)$, then there is some value $b, 0 < b < 1,$ for which

$(y,q) = (x,z; bp, (1-b)r)$.

Axiom 3 (Scale)

For $0 < a < 1, (x, ap) \preceq (y, aq)$ if and only if $(x,p) \preceq (y,q)$.

Axiom 4 (Independence)

Given axiom 3, for $(x, z; ap, (1-a)r) \preceq (y, z; aq, (1-a)r)$ if and only if $(x,p) \preceq (y,q)$.

Theorem :
The individual's preferences, under the N–M axioms (1) to (4), can be defined over prospects (x, p) by a utility function of the form $pv(x)$, where $pv(x) = \sum p_i v(x_i)$.

Proof :
Let M and N be such that $(M, 1) \succ (x_i, 1)$ for all x_i and $(N, 1) \prec (x_i, 1)$ for all x_i. v_M is the probability of getting M and v_N is the probability of getting N. For any prospect $(x_i, 1)$, since $N \prec x_i \prec M$, there exists, by N–M axioms 1 and 2, a prospect $(M, N; v_{Mi}, v_{Ni})$ such that:

$$(x_i, 1) = (M, N; v_{Mi}, v_{Ni})$$

where v_{Mi} and v_{Ni} are functions of x_i. Hence, by axioms 3 and 4,

$$\begin{aligned}
&(x_1, x_2, \ldots, x_n; p_1, p_2, \ldots, p_n) \\
&\quad = \left[(M, N; v_{M1}, v_{N1}), (M, N; v_{M2}, v_{N2}), \ldots, (M, N; v_{Mn}, v_{Nn}); p_1, p_2, \ldots, p_n\right] \\
&\quad = (M, N; v_{M1}p_1 + v_{M2}p_2 + \cdots + v_{Mn}p_n, v_{N1}p_1 + v_{N2}p_2 + \cdots + v_{Nn}p_n)
\end{aligned}$$

where $\Sigma_i v_{Ni} p_i = 1 - \Sigma_i v_{Mi} p_i$, so that $\sum_i v_{Mi} p_i$ can be treated as a utility index of the prospect (x, p), with M and N representing two degrees of freedom. $v_{Mi} = v_M(x_i)$. Hence, the N–M utility index of the prospect (x, p) is $\mathrm{E} p_i v(x_i)$ or, more compactly, $pv(x)$. (*Q.E.D.*)

Axioms 1 and 2 provide the basis for an ordinal utility function, while axioms 3 and 4 lend the utility function its N–M cardinality. Note that there are numerous empirical counterexamples in the literature, from both experimental economics and clinical psychology, showing the violation of one or more of these axioms by both ordinary individuals and their committed exponents. The focus of objections is often to axiom 4, which embodies axiom 3. However, the problem often lies with axiom 3, which leads to the linearity of the expected utility function in probabilities.[31] It is intuitively doubtful whether economic choices follow such linearity, especially as we approach the limits of 0 and 1 for probabilities. Such linearity of utility in probabilities limits the attitude to risk – for instance, the degree of risk aversion – to the slope of the *utility function under certainty*, as can be seen from the absolute and relative measures of risk aversion, as well as from the Friedman and Savage (1948) arguments, which were used in this chapter to derive the attitude to risk solely from the slope of the utility function under certainty. Consequently, in the expected utility hypothesis, risk aversion cannot arise from the potential nonlinearity of the utility function under uncertainty in probabilities, which is quite likely to occur in economic choices.

Appendix 2

Opportunity locus for two risky assets

Suppose the market offers only the two risky assets X_1 and X_2, with $\sigma_1, \sigma_2 > 0$. Define $x_1^* = x_1/W$ and $x_2^* = x_2/W = (1 - x_1^*)$, where x_1^* is the proportion of wealth held in the first asset. This changes the budget constraint to:

$$x_1^* + x_2^* = 1 \tag{32}$$

The expected return μ to the portfolio is now given by:

$$\begin{aligned}\mu &= \mu_1 x_1^* + \mu_2(1 - x_1^*) \\ &= \mu_2 + x_1^*(\mu_1 - \mu_2) \\ &= \mu_2 + (\sigma_2 - \sigma_2)(\mu_1 - \mu_2)/(\sigma_1 - \sigma_2) + x_1^*(\sigma_1 - \sigma_2)(\mu_1 - \mu_2)/(\sigma_1 - \sigma_2)\end{aligned}$$

Let $k_1 = (\mu_1 - \mu_2)/(\sigma_1 - \sigma_2)$, so that:

$$\begin{aligned}\mu &= \mu_2 + (\sigma_2 - \sigma_2)k_1 + (\sigma_1 - \sigma_2)k_1 x_1^* \\ &= (\mu_2 - \sigma_2 k_1) + \sigma_2 k_1 + \sigma_1 k_1 x_1^* - \sigma_2 k_1 x_2^*\end{aligned} \tag{33}$$

Now let $k_2 = \mu_2 - \sigma_2 k_1$, with the result that:

$$\mu = k_2 + k_1\{\sigma_1 x_1^* + \sigma_2(1 - x_1^*)\} \tag{34}$$

This is a general result for the expected return on the two-asset portfolio.

31 This linearity is clearly evident if the prospect has only two outcomes x_1 and 0, so that it is $(x_1, 0; p_1, (1 - p_1))$.

(i) Derivation of the opportunity locus for perfect positive correlation ($\rho_{12} = 1$)

For $\rho_{12} = +1$, we have:

$$\sigma = \sigma_1 x_1^* + \sigma_2 x_2^* \tag{35}$$

Substituting (35) in (34), we get:

$$\mu = k_2 + k_1 \sigma \tag{36}$$

where, as mentioned earlier, k_1 and k_2 are constants for given values of the means and standard deviations of the assets. Therefore, in the case of perfect positive correlation between the two assets in the portfolio, the relationship between the portfolio's expected return μ and its standard deviation σ is linear, as shown by the line AB in Figure 5.4a.

For the whole portfolio consisting only of the first asset, we have $x_1^* = 1$ and $(1 - x_1^*) = 0$, so that the preceding equations imply that:

$$\mu = \mu_1 = k_2 + k_1 \sigma_1$$

which, along with $\sigma = \sigma_1$, provides the point A in Figure 5.4a.

For the whole portfolio consisting only of the second asset, we have $x_1^* = 0$ and $x_2^* = (1 - x_1^*) = 1$, so that the preceding equations correspondingly imply that:

$$\mu = \mu_2 = k_2 + k_1 \sigma_2$$

which, along with $\sigma = \sigma_2$, gives the point B in Figure 5.4a.

(ii) Derivation of the opportunity locus for perfect negative correlation ($\rho_{12} = -1$)

In this case,

$$\sigma = \sigma_1 x_1 - \sigma_2 x_2 \tag{37}$$

so that $\sigma = 0$ for $x_1 / x_2 = \sigma_2 / \sigma_1$.

Given (37), we can define a riskless composite asset X_3 for which $\sigma_3 = 0$. It would combine X_1 and X_2 in the proportions given by $x_1 / x_2 = \sigma_2 / \sigma_1$, with $\mu_3 = \{(\mu_1 \sigma_2 + \mu_2 \sigma_1) W\} / (\sigma_1 + \sigma_2)$ and $\sigma_3 = 0$. Now suppose that we have these three assets (X_1, X_2, X_3) in the portfolio. Combinations of only X_1 and X_3 in the portfolio yield $\sigma(X_1, X_3) = 0$ and $\mu(X_1, X_3) = \mu_1 x_1 + \mu_3 x_3$, with the linear opportunity locus AS in Figure 5.4b.[32] Similarly, combinations of only X_2 and X_3 in the portfolio yield $\sigma(X_2, X_3) = 0$ and $\mu(X_2, X_3) = \mu_2 x_2 + \mu_3 x_3$, with the linear opportunity locus SB in Figure 5.4b.

32 σ in $\sigma(X_1, X_3)$ and μ in $\mu(X_1, X_3)$ are being used as functional symbols, so their arguments can clearly indicate which assets are in the portfolio. However, σ and μ retain their definitions as being respectively the standard deviation and the expected return to the portfolio.

In terms of the proportions x_1^* and $x_2^*(=1-x_1^*)$ of the portfolio invested in the two assets with $\rho_{12}=-1$, the standard deviation σ of the portfolio is given by:

$$\begin{aligned}\sigma^2 &= \sigma_1^2 x_1^{*2} - 2\sigma_1\sigma_2 x_1^* x_2^* + \sigma_2^2 x_2^{*2} \\ &= \sigma_1 x_1^* - \sigma_2(1-x_1^*)\end{aligned}$$

so that, for $\sigma=0$, we have:

$$x_1^* = \sigma_2/(\sigma_1+\sigma_2) \tag{38}$$

Substituting (38) in (34), we get, for the portfolio with $\sigma=0$,

$$\mu = c\left[\frac{2\sigma_1\sigma_2}{\sigma_1+\sigma_2} - \sigma_2\right] + \mu_2$$

where

$$c = \frac{\mu_1-\mu_2}{\sigma_1-\sigma_2}$$

This value of μ and $\sigma=0$ specify the point S in Figure 5.4b. Corresponding to this point, we can define a third asset X_3 such that $\sigma_3=0$, and

$$\begin{aligned}\mu_3 &= c\left[\frac{2\sigma_1\sigma_2}{\sigma_1+\sigma_2} - \sigma_2\right] + \mu_2 \\ &= \frac{\mu_1\sigma_2 + \mu_2\sigma_1}{\sigma_1+\sigma_2}\end{aligned}$$

In Figure 5.4b, a portfolio consisting only of X_3 is represented by the point S. Combinations of X_1 and X_3 only in the portfolio would yield the opportunity locus AS and combinations of X_2 and X_3 only would yield the opportunity locus SB. ASB represents the opportunity locus given by the combinations of X_1 and X_2.

Opportunity locus for $-1<\rho_{12}<+1$

The values of μ and σ for the two-asset case are given by:

$$\mu = \mu_1 x_1^* + \mu_2(1-x_1^*) \tag{39}$$

$$\sigma^2 = \sigma_1^2 x_1^{*2} - 2\rho_{12}\sigma_1\sigma_2 x_1^*(1-x_1^*) + \sigma_2^2(1-x_1^*)^2 \tag{40}$$

In the case where both assets entail some risk and ρ_{12} lies between –1 and +1, solving (39) for x_1^* and substituting in (40) gives a non-linear expression for σ, so that the opportunity locus in the (μ, σ) diagram will also be non-linear. It will lie in the area enclosed by ASB in Figure 5.4c. The closer ρ_{12} is to +1, the closer will this locus be to the line AB. When ρ_{12} is less than +1, there exist economies in risk from holding a diversified portfolio and the opportunity locus moves towards ASB.[33]

33 See Note 19 earlier in this chapter.

Summary of critical conclusions

- The speculative demand for money is analyzed for money balances as an asset when the yields on other assets are uncertain.
- Portfolio selection analysis uses the von Neumann–Morgenstern utility function, with maximization of the expected utility of terminal wealth.
- Risk aversion requires a decreasing marginal utility of wealth.
- The functional form of the utility function with constant absolute risk aversion is analytically convenient but has implausible implications for the wealth elasticity of money demand.
- The functional form of the utility function with constant relative risk aversion is analytically convenient and implies a unit wealth elasticity of money demand.
- The speculative demand for M1 and even for M2 may be zero in the modern financially developed economy with alternative assets that are riskless but have higher yields.

Review and discussion questions

1. Compare and contrast Keynes's theory of the speculative demand for money with Tobin's portfolio selection theory utilizing the expected utility hypothesis.
2. Does Tobin's theory imply the potential for volatility that Keynes attributed to the speculative demand for money?
3. "The more volatile are the returns on bonds and stocks, the greater is the demand for money." Can you derive this proposition from Tobin's liquidity preference model? Does it apply to interest-bearing as well as non-interest-bearing money?
4. If the speculative demand for M1 is zero in the modern financially developed economy, is it also zero for some of the broader money aggregates? If not, what are the appropriate scale determinants of the speculative demand for M1 and the broader money aggregates?
5. Assuming a riskless asset called money and two risky assets, analyze the individual's asset demand for money. What will be the general form of the money demand function?

   Further, assuming that the two assets have perfectly negatively correlated returns, derive the implied demand function for money. Use diagrams for your answer.
6. Assume that there are only two assets, money with $\mu_m = 0$ and $\sigma_m = 0$ and bonds with $\mu_b, \sigma_b > 0$, and that the individual has a CARA utility function, so that he maximizes:

   $$EU(W) = \mu_t - \tfrac{1}{2}\gamma_t \sigma_t^2$$

   Now assume that γ fluctuates such that $\gamma_t = \gamma_0 + \eta_t$ and $\gamma_{t+1} = \gamma_0 - \eta_t$. Derive the individual's speculative demand functions M_t^{sp} and M_{t+1}^{sp}.
7. Again, assume that there are only two assets, money with $\mu_m = 0$ and $\sigma_m = 0$ and bonds with $\mu_b, \sigma_b > 0$, and that the individual has a CARA utility function, so that he maximizes:

   $$EU(W) = \mu_t - \tfrac{1}{2}\gamma_t \sigma_t^2$$

   Now assume that σ fluctuates such that $\sigma_t = \sigma_0 + \varepsilon_t$ and $\sigma_{t+1} = \sigma_0 - \varepsilon_t$. Derive the individual's speculative demand functions M_t^{sp} and M_{t+1}^{sp}.
8. In the preceding two questions, what are likely to be the determinants of η and ε in your economy? How volatile are these shifts likely to be over the business cycle?

9. Use your answers to the above questions to discuss Keynes's assertion on the high volatility of the speculative demand for money. Is this assertion still valid? Discuss.
10. Suppose that the individual has the quadratic utility function:

$$U(W) = a + bW + cW^2$$

where W is wealth.

i. Derive the restrictions on a, b, c for a risk averter.
ii. Derive the expected utility function in terms of μ and σ.
iii. Given the plausible assumptions on the utility of wealth, in what range of W is this utility function relevant?

11. Consider a two-asset model with money paying the positive given interest rate R_m and the bonds paying a return R which has the expected value μ_b and standard deviation σ_b. Show diagrammatically the effects of the following for the proportions held of the two assets:

i. The government imposes a tax on the excess return $(R - R_m)$ on bonds, with a corresponding refund if the return is negative.
ii. The government imposes a tax on a positive excess return $(R - R_m)$ on bonds, but without any refund if the return is negative.
iii. Show the effects for the above cases if $R_m = 0$.

12. Consider a two-asset model with money paying the fixed interest rate R_m and the bonds paying the return R with expected value μ_b and standard deviation σ_b. The government imposes a tax on the excess return $(R - R_m)$ on bonds, with a corresponding refund if the return is negative. What will be the effects of the following on bond purchases?

i. The tax revenues are not returned to the investors.
ii. The tax revenues are returned to the investors as a lump-sum transfer.

13. Does the existence of a speculative demand component increase or decrease the interest elasticity of the overall demand for money? When would broader monetary aggregates have higher interest elasticities than narrower ones, especially M1?
14. Evaluate the usefulness and defects of CARA and CRRA utility functions for deriving the demand functions for monetary aggregates. Discuss the likely validity of their implications for elasticity of demand for M1 and M2 with respect to wealth. What are their implications for the elasticity of M1 and M2 with respect to current income?
15. Does the existence of a speculative demand component increase or decrease the income elasticity of the overall demand for money? When would broader monetary aggregates have higher income elasticities than narrower ones, especially M1?
16. "The theory of portfolio choice has little to do with the demand for money in the modern economy." Discuss.
17. "Liquidity preference as behavior towards risk is a demand for short-term securities – not money." Present Tobin's analysis of the demand for money.

Use Tobin's analysis, or any other, to show the conditions under which the above (quoted) statement will apply. How will the accuracy of this statement modify the demand for money?

References

Arrow, K.J. "The theory of risk aversion." In K.J. Arrow, *Essays in the Theory of Risk-Bearing*. Chicago: Markham, 1971.

Chang, W.W., Hamberg, D. and Hirata, J. "Liquidity preference as behavior toward risk is a demand for short-term securities – not money." *American Economic Review*, 73, 1983, pp. 420–7.

Cuthbertson, K. *The Supply and Demand for Money.* London: Blackwell, 1985.

Friedman, M., and Savage, L.J. "The utility analysis of choices involving risk." *Journal of Political Economy*, 56, 1948, pp. 279–304.

Handa, J. "Decisions under imperfect knowledge: the certainty equivalence theory as an alternative to the Von Neumann–Morgenstern theory of uncertainty." *Erkenntnis,* 20, 1983, pp. 295–328.

Keynes, J.M. *The General Theory of Employment, Interest and Money*. London: Macmillan, 1936, Chs. 13, 15.

Tobin, J. "Liquidity preference as behavior towards risk." *Review of Economic Studies*, 25, 1958, pp. 65–86.

Tobin, J. "The theory of portfolio selection." In F.H. Hahn and F.P.R. Brechling, eds, *The Theory of Interest Rates*. London: Macmillan, 1965.

6 Precautionary and buffer stock demand for money

Keynes referred to the precautionary motive for holding money but did not present any analysis of it. This demand arises from uncertainty of income and the need for expenditures.

This chapter presents the extension of the transactions demand and speculative demand analyses to the precautionary money demand. An additional source of the demand for money is buffer stock, which arises because of lags in the adjustment of income, commodities and bonds.

Key concepts introduced in this chapter

- Uncertainty of consumption expenditures and income
- Precautionary demand for money
- Buffer stock demand for money
- Overdrafts and stand-by credit arrangements
- The dependence of the demand for money on money supply changes

Neither the analysis of the transactions demand for money in Chapter 4 nor that of the speculative demand for money in Chapter 5 dealt with the uncertainty of income or of the need for expenditures in the future. Such uncertainty is pervasive in the economy, and the individual can respond to it through precautionary saving, some or all of which could be held in the form of precautionary money balances.

Precautionary saving is that part of income that is saved because of the uncertainty of future income and consumption needs. It would be zero if the future values of these variables were fully known. Precautionary wealth or savings are similarly the part of wealth held due to such uncertainty. Such wealth may be held in a variety of assets, one of which is money. Money balances held for such a reason constitute the precautionary demand for money. Saving and precautionary money balances are thus different concepts, with saving being the means of carrying purchasing power from one period to the next and precautionary money balances being the means of paying for unexpected expenditures within any given period.

Precautionary wealth is clearly affected by the economic and financial environment, as well as by the individual's own personal circumstances. The economic environment – which includes the possibility of being laid off or, if unemployed, of finding a job, the growth of incomes, the social welfare net, etc. – is one of the determinants of the uncertainty of the individual's future income. The economy's financial structure provides for such devices as

credit cards, overdrafts, trade credit, etc., which allow for payments for sudden expenditures to be postponed and reduce the need for the precautionary holding of assets. The individual's personal circumstances affect his expenditure needs, the timing of expenditures and the possibility of delaying them, or temporarily meeting them through the use of credit cards, overdrafts, etc. The precautionary demand for money depends upon the above factors and the relative liquidity and transactions costs of the various assets that can function as precautionary wealth.

Since the focus of the speculative demand for money is on the uncertainty of the yields on the various assets, the precautionary demand for money is, for simplicity, analyzed under the assumption that these yields are known – and therefore are not uncertain. Given this assumption, the analysis of the precautionary demand for money becomes an extension of the inventory analysis of the transactions demand to the case of uncertainty of the amount and timing of income receipts and expenditures. This uncertainty of income is captured through the moments of the income distribution, with the analysis assuming a normal distribution and therefore considering only the mean and variance of income during the period.

The inventory analysis of transactions demand in Chapter 4 implied the general version of the demand function as:

$$m^{\text{trd}} = m^{\text{trd}}(b, R, y) \tag{1}$$

where:

m^{trd} = transactions demand for real balances
b = real brokerage cost
R = nominal interest rate
y = real income/expenditures.

Under uncertainty, assuming a normal distribution of income, y is a function of its mean value and standard deviation. Hence, the modification of (1) to the case of precautionary demand (subsuming within it the transactions demand) for money, would be:

$$m^{\text{prd}} = m^{\text{prd}}(b, R, \mu_y, \sigma_y) \tag{2}$$

where:

m^{prd} = precautionary demand for money
μ_y = mathematical expectation of income
σ_y = standard deviation of income.

In addition, under uncertainty, the degree of risk aversion and the available mechanisms and substitutes for coping with such uncertainty would also be among the relevant determinants of money demand. That is, (2) needs to be modified to:

$$m^{\text{prd}} = m^{\text{prd}}(b, R, \mu_y, \sigma_y, \rho, \Omega) \tag{3}$$

where:

ρ = degree of risk aversion
Ω = substitutes for precautionary money balances.

Among the components of Ω would be credit cards, overdrafts at banks, trade credit, etc. In the limiting case in which the individual can pay for any precautionary needs by credit cards

and pay the credit card balances on the date he receives income, his precautionary demand for money would be zero.

Note that, in (3), if the individual were risk indifferent, $\rho = 0$ and (3) would simplify to:

$$m^{\text{prd}} = m^{\text{prd}}(b, R, \mu_y, \Omega) \tag{4}$$

The preceding arguments imply a *unique* value for the precautionary demand for money for given values of its determinants. Such models are presented in Sections 6.1 to 6.3. Somewhat different from these models are those known as buffer stock models. In a *buffer stock model*, money is held as a "buffer" or fallback because money has lower transactions costs than other assets, so that the receipt of income can be held in the form of money until a sufficiently large amount has accumulated for it to be worthwhile to adjust other assets or income–expenditure flows. The actual holdings of money would therefore exhibit "short-run" fluctuations, implying that the short-run money-demand function and velocity would be unstable, though within a specific range. There are two common patterns of such short – run fluctuations. One of these is fluctuation around a long–run desired level; the other is fluctuation within a band whose upper and lower limits are specified by longer-term factors. Milbourne (1988) provides a survey of buffer stock models.

This chapter presents, in Sections 6.1 to 6.3, precautionary demand models that use the contributions of Whalen (1966) and Sprenkle and Miller (1980), which imply determinate levels of the precautionary demand for money rather than fluctuations around a desired level or in an optimal range. Sections 6.4 to 6.7 present some of the buffer stock models and empirical findings.

The economic agent can be the individual/household or firm, though some of the contributions in the literature refer specifically to the firm, some to the individual and some to the (economic) agent. We will use the terms individual, firm and agent interchangeably in the following, with the understanding that the analysis is to be applied as appropriate.

6.1 An extension of the transactions demand model to precautionary demand

The following analysis of the precautionary demand for money is based on Whalen (1966). Assume, as in the inventory model of transactions demand, that the individual has a choice between holding money or bonds. Money is perfectly liquid and does not pay any interest. Bonds are illiquid and pay interest at the rate r. There is a brokerage cost of converting from money to bonds and vice versa. Further, as an item additional to the ones in the transactions demand model of Chapter 4, selling bonds at short notice to obtain money for unexpected transactions or having to postpone such transactions imposes an additional "penalty" cost. Therefore, there are now three components of the cost of financing transactions: brokerage costs, interest income foregone and penalty costs. As in Chapter 4, the individual is assumed to withdraw an amount W from bonds at evenly spaced intervals.

The cost function associated with money usage is:

$$C = RM + B_0 Y/W + \beta p(N > M) \tag{5}$$

where:

C = nominal cost of holding precautionary (including transactions) balances
M = money balances held
B_0 = nominal brokerage cost per withdrawal
Y = total (uncertain) nominal income/expenditures
W = amount withdrawn each time from interest-bearing bonds
N = net payments (expenditures less receipts)
$p(N > M)$ = probability of $N > M$
β = nominal penalty cost of shortfall in money balances.

Since the individual has an uncertain pattern of receipts and payments and needs to pay for any purchases in money, he suffers a loss ("penalty") whenever he is short of money to make an intended purchase. This loss can be that of having to unexpectedly sell "bonds" to get the required money balances or having to postpone the purchase until he has enough money, so that this loss can have monetary and non-monetary components. With p as the probability of $N > M$, (5) specifies the penalty cost of having inadequate balances by $\beta p(N > M)$.

Suppose that the individual holds money balances M equal to $k\sigma$ where σ is the standard deviation of net payments N, so that:

$$M = k\sigma \tag{6}$$

We need to know the probability $p(N > k\sigma)$ that the net payments N will exceed money holdings $k\sigma$ so that the penalty will be incurred. By Chebyscheff's inequality, the probability p that a variable N will deviate from its mean – which is zero under our assumptions – by more than k times its standard deviation σ is specified by $p(-k\sigma > N > k\sigma) \leq 1/k^2$. Therefore $p(N > M)$, where M equals $k\sigma$, is:

$$p(N > M) \leq 1/k^2 \qquad k \geq 1^{1} \tag{7}$$

where, from (6),

$$k = M/\sigma \tag{8}$$

Assume that the individual is sufficiently risk averse to base his money holdings on the maximum value of $p(N > M)$. In this case,

$$p(N > M) = 1/(M/\sigma)^2 = \sigma^2/M^2 \tag{9}$$

Equations (5) and (9) imply that:

$$C = RM + B_0 Y/W + \beta\sigma^2/M^2$$

Therefore, since $M = \frac{1}{2}W$, as in Baumol's analysis,

$$C = RM + \tfrac{1}{2}B_0 Y/M + \beta\sigma^2/M^2 \tag{10}$$

1 This assumption is being made to ensure that the maximum value of $p(N > M)$ is less than or equal to one.

Note that the first two terms of the above equation are as in Baumol's analysis. The third term arises because of the uncertainty of expenditures. To minimize the cost of holding money, set the partial derivative of C with respect to M equal to zero, as in:

$$\partial C/\partial M = R - \tfrac{1}{2}B_0 Y/M^2 - 2\beta\sigma^2/M^3 = 0 \tag{11}$$

Multiplying by M^3,

$$RM^3 - \tfrac{1}{2}B_0 MY - 2\beta\sigma^2 = 0 \tag{12}$$

Equation (12) specifies a cubic function in M and is, in general, difficult to solve. To simplify further, we can make one of two possible simplifying assumptions.

(i) If there is no penalty cost to a shortfall in money holdings, $\beta = 0$, while if there is no risk of such a shortfall, $\sigma = 0$. For either $\beta = 0$ or $\sigma = 0$, (12) reduces to Baumol's demand function for transactions balances, given in Chapter 4. This was:

$$M^{\text{tr}} = (\tfrac{1}{2}B_0)^{\frac{1}{2}} Y^{\frac{1}{2}} R^{-\frac{1}{2}}$$

However, the simplifying assumption $\beta = 0$ or $\sigma = 0$ eliminates the precautionary demand elements, so that for precautionary demand analysis we opt for the following simplification.

(ii) Assume that while β, $\sigma > 0$, the brokerage cost is zero, so that $B_0 = 0$.[2] Making this assumption:

$$RM^{\text{pr}3} - 2\beta\sigma^2 = 0$$

so that:

$$M^{\text{pr}} = (2\beta)^{1/3} R^{-1/3} (\sigma^2)^{1/3} \tag{13}$$

The particular insight of (13) is that the precautionary demand for money will depend upon the variance σ^2 of net income and not necessarily on the level of income itself. By comparison, the transactions demand for money in Baumol's analysis depended upon income or, in the present context, on the expected level of income. In (13), the average level of income/expenditures Y has dropped out of the money demand function because of the elimination of the brokerage cost term ($\tfrac{1}{2}B_0 Y/M$) in the simplification in going from (12) to (13). This simplification, therefore, eliminates the transactions demand for money, which is related to the level of expenditures, so that (13) should be taken as specifying the precautionary demand – *exclusive* of the transactions demand – for money. In keeping with this, the superscript pr has been added to the money symbol in (13).

From (13), the interest elasticity of the precautionary demand is $-1/3$, not $-1/2$.

Now assume that the time pattern of receipts and payments during the period does not change but their amounts vary proportionately with the total expenditures Y over the period.

2 Note that making the assumption that $B_0 = 0$ in Baumol's model would imply that the transactions demand for money is zero. Hence, this assumption eliminates the analysis of the transactions demand from the present model.

In this case, for a normal distribution of net payments, the variance of receipts and payments will increase proportionately with Y^2. Let this be represented by:

$$\sigma^2 = \alpha Y^2 \tag{14}$$

where α is a constant whose value depends on the given time frequency of receipts and payments. From (13) and (14), we get:

$$M^{\text{pr}} = (2\alpha\beta)^{1/3} R^{-1/3} Y^{2/3} \tag{15}$$

so that the elasticity of precautionary balances with respect to the amount of income/expenditures will be 2/3.

However, if the amounts of the payments and receipts do not change but their number increases proportionately with Y so that they become more frequent as Y increases, σ^2 will change proportionately with Y, such that $\sigma^2 = \alpha' Y$. The demand for money in this case would be:

$$M^{\text{pr}} = (2\alpha'\beta)^{1/3} R^{-1/3} Y^{1/3} \tag{16}$$

so that the income elasticity of precautionary balances is now only 1/3.

Since expenditures can change in the real world in either of the two ways envisaged in (15) and (16) or in other ways, the implied income elasticity of the precautionary money demand lies in the range from 1/3 to 2/3, depending upon how income and expenditures change. Further, note that since the transactions demand was dropped out of the model in simplifying from (12) to (13), (15) and (16) do not provide any information on the transactions demand elasticities, so that these equations specify only the demand for precautionary balances. If we had been able to solve (12) for M, such a solution would have provided a combined money demand for both transactions and precautionary purposes, but there is no guarantee that this solution would have an income elasticity of 1/2, 1/3 or 2/3. Further, even for the precautionary demand alone, as in (15) and (16), the actual elasticity will not necessarily be 1/3 or 2/3 if the distribution of net payments is not normal or if the time pattern or the amount of individual transactions during the period both change simultaneously.[3] Empirically estimated *real income elasticities* of the demand for real balances of M1 (currency and demand deposits) in the economy tend to be somewhat below unity, but not as low as 1/3. The income elasticity of 1/3 in (16) is therefore quite unrealistic, which is not surprising since it excludes the transactions demand and also assumes that the amounts of the individual transactions do not change. However, its *interest elasticity* of $(-1/3)$ is closer to the empirically estimated values than its value of $(-1/2)$ in the transactions model.

To examine the elasticity of the demand for precautionary balances with respect to the price level, first note that β is the nominal penalty cost. Set it equal to $\beta^{\text{r}}P$, where β^{r} is the real penalty cost and P is the price level. Also assume that the increase in the price level increases the magnitudes of all receipts and payments proportionately, while leaving their timing unchanged. Hence, with $Y = Py$, where Py is nominal expenditures and y is their real value, $\sigma^2 = \alpha P^2 y^2$, so that (15) becomes:

$$m^{\text{pr}} = M^{\text{pr}}/P = (\alpha\beta^{\text{r}})^{1/3} R^{-1/3} y^{2/3} \tag{17}$$

3 Whalen (1966) also presents in the appendix to his article two other variations of the model presented above.

so that the demand for real precautionary balances is homogeneous of degree zero in the price level. Such homogeneity of degree zero of real balances does not hold in the context of (16), which has a price elasticity of only 2/3 for nominal balances.

6.2 Precautionary demand for money with overdrafts

The preceding model from Whalen assumes that the individual does not have automatic access to overdrafts. This is often not the case for large – and sometimes even for small – firms. It is also not the case for many individuals who have arranged overdraft/credit facilities with their banks or who can resort to credit cards, whose limits can be treated as overdraft limits. The analysis of this case and its variations in the following is from Sprenkle and Miller (S–M) (1980). These authors analyze three cases, with no-limit overdrafts, with overdraft limits and without overdrafts. These cases can apply to firms as well as households. However, S–M consider the no-limit overdraft case to be especially applicable to large firms and the no-overdraft case to be the most pertinent one for households.

S–M assume that the economic agent – which will be taken to be a firm in this case but could instead be a household – has an overdraft from a bank and wants to minimize the cost of holding precautionary balances. If it holds larger balances than needed, it foregoes the interest rate R from investing the funds in bonds; if it holds inadequate precautionary balances, it has to pay the interest rate ρ on overdrafts but earns the rate R on the balances held in bonds, so that the net loss in using overdrafts is only $(\rho - R)$.[4] We will assume that $\rho > R$. The cost C of holding precautionary balances is therefore:

$$C = R\int_{-\infty}^{A}(A-Z)f(Z)\,\mathrm{d}Z + (\rho - R)\int_{A}^{\infty}(Z-A)f(Z)\,\mathrm{d}Z \tag{18}$$

where:

A = precautionary balances at the beginning of the period
Z = payments less receipts
$f(Z)$ = probability distribution of Z, with $f(\infty) = 0$
R = nominal interest rate on bonds
ρ = nominal interest rate charged on overdrafts.

S–M treat Z as a forecast error with $E(Z) = 0$ so that the payments and receipts over the period are equal. (18) involves deciding on the amount A at the beginning of the period to cover the possibility of overdraft charges – as against the brokerage charges for withdrawals from bonds that would have occurred in the absence of overdrafts, as in Baumol's analysis of the transactions demand for money. Minimizing (18) with respect to A yields:

$$\mathrm{d}C/\mathrm{d}A = R\int_{-\infty}^{A} f(Z)\,\mathrm{d}Z - (\rho - R)\int_{A}^{\infty} f(Z)\,\mathrm{d}Z \tag{19}$$

Designating $F(Z)$ as the cumulative probability distribution of Z, with $F(\infty) = 1$, (19) becomes:

$$\mathrm{d}C/\mathrm{d}A = R - \rho[1 - F(A)] = 0 \tag{20}$$

4 In the presence of overdrafts, $(\rho - R)$ corresponds to Whalen's penalty cost of being caught short of funds to pay for expenditures.

so that

$$F(A^*) = (\rho - R)/\rho \tag{21}$$

where A^* is the optimal value of A and $F(A^*)$ is the cumulative probability that the optimal cash holdings will at least equal the need for cash.

To derive the optimal precautionary holdings, note that the precautionary balances will be zero when actual overdrafts are positive. Therefore, the precautionary balances M^{pr} only equal:

$$M^{\text{pr}} = \int_{-\infty}^{A^*} (A^* - Z) f(Z)\, \mathrm{d}Z \tag{22}$$

Assuming a normal distribution with zero mean:

$$M^{\text{pr}} = A^* F(A^*) + \sigma f(A^*) \tag{23}$$

where σ is the standard deviation of Z. In the case of a normal distribution, the average amount borrowed through overdrafts (O) will be:

$$O = -A^*[1 - F(A^*)] + \sigma f(A^*) \tag{24}$$

From (21) and (23),

$$\partial M^{\text{pr}}/\partial \rho = \left[\left\{1 - F(A^*)\right\} F(A^*)\right]/\rho f(A^*) > 0 \tag{25}$$

$$\partial M^{\text{pr}}/\partial R = -F(A^*)/\rho f(A^*) < 0 \tag{26}$$

so that M^{pr} decreases (and overdraft borrowings increase) when the bond rate R increases and the overdraft rate ρ decreases. These are intuitively plausible results since a rise in R and a decrease in ρ increase the opportunity cost of holding precautionary balances.

There is no easy way to derive the interest elasticity of the precautionary money demand for (23) unless the probability distributions $f(A)$ and $F(A)$ are first specified. What is clear from (23) is that this demand depends upon these distributions and therefore upon their moments. Assuming a normal distribution, this demand will depend upon the mean and variance of expenditures, as in Section 6.1.

Equation (23) specifies the precautionary demand for money and does not include the transactions demand. To compare the above analysis with that of Baumol's transactions demand under certainty, now add the assumption of certainty to the present analysis. The assumption for the uncertainty case was that $E(Z) = 0$. In the case of certainty, Z always has a known value. If $Z \leq 0$ (that is, receipts exceed payments every time), C in (18) would be zero and so would be the demand for money derived from it. But if payments exceed receipts at different points during the period, so that $Z > 0$, the individual will prefer to start with enough transactions balances in order to avoid using the more costly overdrafts with $\rho > R$. Hence, under certainty, the precautionary demand for money (exclusive of transactions demand) in this analysis will be zero.

6.3 Precautionary demand for money without overdrafts

If the economic agent is not allowed overdrafts, the total cost consists of the interest lost on not holding bonds and the cost of being overdrawn, which is assumed for the time being to be equal to or less than the cost of having to postpone expenditures, so that:

$$C = R\int_{-\infty}^{A}(A-Z)f(Z)\,\mathrm{d}Z + \beta\int_{A}^{\infty}f(Z)\,\mathrm{d}Z \tag{27}$$

where β is the penalty to being overdrawn and the second integral on the right-hand side is the probability of being overdrawn. Hence, the second term in (27) is the cost of being overdrawn. Minimizing (27) implies that:

$$F(A^*)/f(A^*) = \beta/R \tag{28}$$

Since β is the penalty cost of holding inadequate balances, it can be compared with $(\rho - R)$ in the analysis of Section 6.2, where ρ was the interest charged by the bank on overdraft balances. In the present analysis, if the banks wanted to discourage some customers from being overdrawn they would set ρ and β fairly high, so that for such customers the operative value of β would become the penalty cost of finding funds elsewhere or of postponing expenditures.

The response of M1 to the interest rate R, where R is the cost of holding M1, was derived from (28) by S–M as:

$$\partial \mathrm{M1}^{\mathrm{pr}}/\partial R = \left\{F(A^*)^2\right\}/[R.f(A^*) - \beta f'(A^*)] < 0 \tag{29}^{5}$$

where f' is the partial derivative of f with respect to A^* and money is M1(which excludes the assets on which the interest rate R is paid). The above analysis is very similar to that of Section 6.2, with the penalty rate β corresponding to the overdraft interest charge ρ.

If money is defined very broadly as M3 to include the closest substitutes for M1, and R_s is the interest rate on such substitutes, the S–M derivation showed that:

$$\partial \mathrm{M3}^{\mathrm{pr}}/\partial R_{\mathrm{s}} = \left[F(A^*)\left\{1 - F(A^*)\right\}\right] \Big/ \left[\rho f(A^*) - \beta f'(A^*)\right] > 0 \tag{30}$$

The difference in the signs of the partial derivatives in (29) and (30) occurs because R in (29) is the return on the alternative assets to M1 and therefore part of the cost of holding it, whereas R_{s} in (30) is the return on one of the (short-term) assets in M3.

The above two cases – with a no-limit overdraft and without an overdraft – illustrate the basic nature of the S–M analysis; their analysis of the intermediate case of a binding limit on the overdraft is not presented here. The two analyses imply that, under uncertainty of the timing of receipts and payments, there will be a positive precautionary demand for money. This demand has the general form:

$$M^{\mathrm{pr}} = M^{\mathrm{pr}}(R, \rho, \beta, f(z)) \tag{31}$$

5 For the derivation of (28) and (29) from (27), see Sprenkle and Miller (1980, p. 417).

The elasticity of precautionary balances with respect to the bond rate R is negative. However, it is not possible to derive the interest and income elasticities of M1 and M2 without further specification of the probability distribution of expenditures.[6]

6.4 Buffer stock models

The theoretical analysis of buffer stock models extends the inventory analysis of the transactions demand for money to the case of uncertainty of net payments (payments less receipts), as in the case of the precautionary demand models of Sections 6.1 to 6.3. However, while this precautionary demand analysis has determined an optimal amount of precautionary balances, the buffer stock models allow short-run money balances to fluctuate within a band which has upper and lower limits, also known as thresholds, or fluctuate around a long-run desired money demand.

There are basically two versions of buffer stock models. In one of these, a "policy decision" is made a priori by the individual that cash balances will be allowed to vary within an upper (M_{max}) and a lower limit (M_{min}). This case is depicted in Figure 6.1. When the autonomous – that is, independent of the decision to invest in bonds or disinvest from bonds – net receipts cause the accumulated cash balance to hit the upper limit M_{max}, action is taken to invest a certain amount in other assets, say bonds, thereby reducing cash holdings suddenly by the corresponding amount. Whenever the autonomous net payments deplete the cash reserves sufficiently to reach the minimum permitted level M_{min}, action is taken to rebuild them by selling some of the bonds. This lower limit can be zero or a positive amount, depending upon institutional practices such as minimum balances required by banks, etc. Such buffer stock models with a pre-set band belong to the (Z, z) – with Z as the upper limit and z as the lower one – type of inventory models and are called "*rule models*", where the rule specifies the adjustment made when the money balances hit either of the limits.

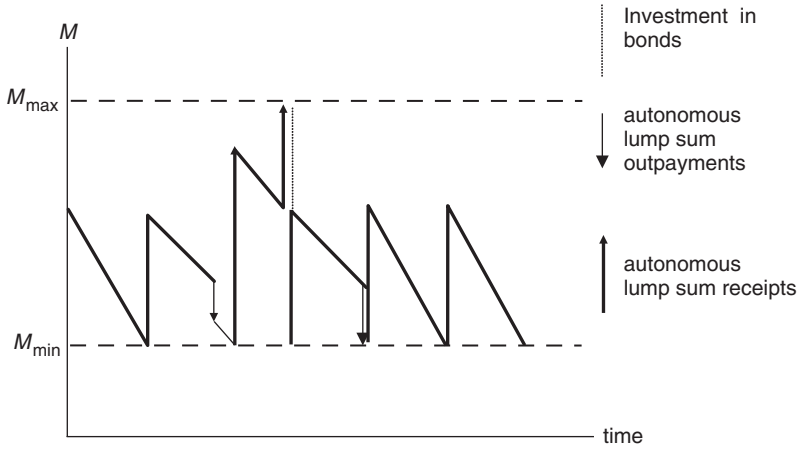


Figure 6.1

6 Sprenkle and Miller (1980) use some numerical examples to provide guidance on the money demand given by the above analysis.

In such rule models, money balances can change because of positive or negative net payments or because of action taken by the agent to reduce them when they reach the upper limit or increase them when they reach the lower limit. The former can be designated as "autonomous" or "exogenous" changes and the latter as "induced" changes in money balances. In the former, the change occurs even though the agent's objective is not to adjust his money holdings. In the latter, the agent's intention is to adjust the money balances since they have moved outside the designated band.

The second type of buffer stock models is called "*smoothing* or *objective models*." In these, the objective is to smooth movements in other variables such as consumption or expenditures and bond holdings. Unexpected increases in income receipts or decreases in payments would be added to money balances acting as the "residual" inventory or temporary abode of purchasing power until adjustments in expenditures and bond holdings can be made. Conversely, unexpected decreases in income receipts or increases in payments would be temporarily accommodated by running down money balances, rather than through an immediate cutback in expenditures or sales of bonds. The reason for thus treating money holdings as a residual repository of purchasing power is that the cost of small and continual adjustments in such balances is assumed to be lower than in either expenditures or payments, or in bond holdings, so that temporarily allowing such balances to change is the optimal strategy. In such smoothing models, actual balances fluctuate around their desired long-run demand, but there are no pre-set upper and lower limits as in the case of the rule models. This case is illustrated in Figure 6.2. Note that the distinction between the autonomous and induced (causes of) changes in money balances applies in both smoothing and rule models.

There can be quite a number of models of each variety. This chapter examines two models of each of the two versions of the buffer stock models. The models presented for the rule version are those by Akerlof and Milbourne (A–M) (1980) and Miller and Orr (M–O) (1966). The models presented for the smoothing version are those of Cuthbertson and Taylor (C–T) (1987) and Kanniainen and Tarkka (K–T) (1986).

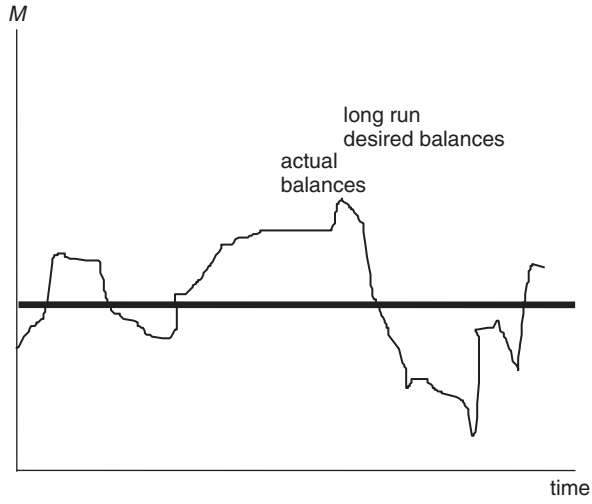


Figure 6.2

6.5 Buffer stock rule models

6.5.1 The rule model of Akerlof and Milbourne

We start the analysis of buffer stock models with the contribution of Akerlof and Milbourne (A–M) (1980). As in Baumol's transactions demand model, A–M[7] assume a lump-sum receipt of Y at the beginning of the period and expenditures of C at a constant rate through the period. However, in a departure from Baumol's model, A–M assume that $C \leq Y$, with saving $S = Y - C$. Saving during the period is added to money balances until the latter reach the set upper limit, at which time action is taken to decrease them to C through their partial investment in bonds, so that they are expected to be exhausted by the next period.

Designate the upper limit as Z and the lower one as z, with the latter taken to be zero for simplification. The agent wishes to start each period with the amount C, which will therefore be the desired amount at the beginning of each period. The actual amounts held at the beginning of the ith period equal $C + \mathrm{i}S$, as long as $(C + \mathrm{i}S) \leq Z$. If the upper limit is reached after n periods, we have:

$$C + nS \geq Z \geq C + (n-1)S \tag{32}$$

so that:

$$n \geq (Z - C)/S \geq (n-1) \tag{33}$$

Hence, the maximum and minimum values of n are:

$$n_{\max} = [(Z - C)/S] + 1 \tag{34}$$

$$n_{\min} = [(Z - C)/S] \tag{35}$$

The amount C is spent evenly during the period so that the average amount of cash balances corresponding to it is $C/2$. When saving S is added to money balances in a period, this amount is held from the beginning of the period to its end, so that the average cash balances corresponding to it are S. Therefore, the sequence of money balances at the end of each period is:

$$\{C/2, C/2 + S, C/2 + 2S, \ldots, C/2 + (n-1)S\} \tag{36}$$

which equals:

$$(C/2)\{1, 1 + S, 1 + 2S, \ldots, 1 + (n-1)S\}$$

7 The model presented below is only the basic one from Akerlof and Milbourne. For its more complex forms and for numerical illustrations, see the original article.

Let n be the number of periods before an induced transfer takes place. Then, over n periods, the average balance *between induced adjustments* is:

$$M^{\mathrm{d}} = \frac{1}{n}\left[\frac{C}{2} + \left(\frac{C}{2} + S\right) + \left(\frac{C}{2} + 2S\right) + \cdots + \left\{\frac{C}{2} + (n-1)S\right\}\right]$$
$$= \frac{C}{2} + \frac{S(n-1)}{2} \tag{37}$$

Using (34) and (35) to eliminate n in (37) implies the minimum and maximum values of money balances as:

$$M_{\max} = Z/2 \tag{38}$$

$$M_{\min} = (Z - S)/2 \tag{39}$$

Hence, the average of the money balances held as a buffer stock, designated as $\boldsymbol{M}^{\mathrm{b}}$, is:

$$\begin{aligned} \boldsymbol{M}^{\mathrm{b}} &= (1/2)(M_{\max} + M_{\min}) \\ &= Z/2 - S/4 \end{aligned} \tag{40}$$

so that:

$$\partial \boldsymbol{M}^{\mathrm{b}}/\partial Y \cong -(1/4)(\partial S/\partial Y) < 0 \tag{41}$$

where $\partial S/\partial Y$ is the marginal propensity to save, which is positive. Hence, $\partial \boldsymbol{M}^{\mathrm{b}}/\partial Y$ is *negative*. This is a surprising result. Its intuitive explanation is that, as income rises, the upper threshold is reached more quickly, so that the interval before the money holdings are run down by an induced adjustment is shortened. As a result, the richer agents review their money and bond holdings more often than those with lower incomes, *ceteris paribus*, and will hold less balances on average.[8]

However, since the limits z and Z were assumed to be exogenously specified in the above model, the impact of increases in Y on them is not incorporated in (40). Transactions demand analysis implies that both of these limits would be positive functions of the level of expenditures.[9] Therefore, the impact of a rise in income would have a positive and a negative component, with the net impact being of indeterminate sign unless a fuller model were specified. Another limitation of the above model is that it does not distinguish between the expected increases in income and the unexpected ones. The (z, Z) concept is more appropriate to the latter than to the former.

Akerlof and Milbourne extended their preceding model to the case of the uncertainty of net payments by assuming that the agent buys and pays for durable goods at uncertain times, with p as the probability of making such a purchase. For this case, the A–M result, under their

8 This result does not hold if the upper limit X is defined relative to income Y as xY.

9 If we follow the pattern of Baumol's transactions analysis, suitably adapted to the present context, z and Z will be non-linear functions of Y.

simplifying assumptions that include $S = sY$, where s is the constant average propensity to save, was that:

$$\boldsymbol{M}^{\mathrm{b}} \approx Z/2 - s(1+p)(Y/4) < 0 \tag{42}$$

where p is the probability of payments (for a durable goods purchase). (42) implies that:

$$\partial \boldsymbol{M}^{\mathrm{b}}/\partial Y = -(s/4)[1 + p + Yp'(Y)] \tag{43}$$

where $p' = \partial p/\partial Y$. Assuming p' to be positive – that is, the probability of buying durable goods increases with income – (43) implies that the income elasticity of money balances is again *negative*.

The A–M model was meant to encompass both the transactions and precautionary demands. The agent's desire to finance an amount of transactions C out of income receipts Y creates his transactions demand, while the uncertainty of payments adds an additional precautionary demand. However, this framework does not properly capture the transactions demand since it ignores the dependence of (z, Z) on total expenditures and does not make a distinction between expected and unexpected changes in income. Its implication of a negative income elasticity of money demand must therefore be accepted as reflecting the influence of saving, and especially unexpected saving, on money demand, with consumption – and hence permanent income – being held constant. Some of the ideas behind these criticisms will become clearer from the theoretical and empirical models presented below.

6.5.2 The rule model of Miller and Orr

Miller and Orr (M–O) (1966) assumed that net receipts – which would be net payments for a negative value – of x at any moment follow a random walk with a zero mean over each period (e.g. a "day").[10] Assume that in any time interval (e.g. an "hour") equaling $1/t(t = 24)$, x is generated as a sequence of independent Bernoulli trials. The individual believes with a subjective probability of p that he will have net receipts of x during each time interval (hour) or net payments of x with a probability of $(1 - p)$, so that, over an hour, the probability of an increase in money holdings by x is p and that of a decrease by x is $(1 - p)$.

Cash holdings over a decision period of T periods will have a mean and standard deviation given by:

$$\mu_T = Ttx(p - q) \tag{44}$$

$$\sigma_T^2 = 4Ttpqx^2 \tag{45}$$

where:

x = net receipts per hour
μ_T = mean cash holdings over T periods
σ_T = standard deviation of cash holdings over T periods
p = probability of positive net payments
q = probability of negative net payments $(= 1 - p)$
t = number of sub-intervals (hours) in each period (day)
T = number of time periods up to the planning horizon.

10 Under this assumption, the pattern of net payments will possess stationarity and serial independence.

For simplification, p was assumed by M–O to be $1/2$,[11] so that:

$$\mu_T = 0 \tag{46}$$

$$\sigma_T^2 = Ttx^2 \tag{47}$$

Since the variance of changes in cash holdings (σ_T^2) during T periods is Ttx^2, the variance σ^2 of daily changes in balances (over the day) is:

$$\sigma^2 = \sigma_T^2/T = tx^2 \tag{48}$$

The cost of holding and varying cash has two components: the interest cost of holding cash rather than bonds, and the brokerage cost of making deliberate changes in cash holdings. The per period (daily) expected cost is:

$$E(TC) = B_0E(N)/T + RM \tag{49}$$

where:

$E(TC)$ = expected cost per period of holding and managing cash
M = average daily cash balance
$E(N)$ = expected number of transactions between money and bonds over T periods
B_0 = brokerage cost per transaction
R = interest rate per period (day) on bonds.

The firm is taken to minimize (49) with respect to the upper limit Z and the lower limit z on cash balances. M–O show that under certain specific assumptions:

$$E(N)/T \to 1/D(z, Z)$$

where D is the mean of the time intervals separating portfolio transfers between bonds and cash, and that:

$$D(z, Z) = z(Z - z)/tx^2 \tag{50}$$

Further, M–O show that the steady–state distribution of money balances during the day has a discrete triangular distribution for $p = 1/2$, so that the average balances M are given by:

$$M = (Z + z)/3 \tag{51}$$

Hence, using the maximum value of $E(N)/T$, (51) can be restated as:

$$E(TC) = B_0\ tx^2/zw + R(w + 2z)/3 \tag{52}$$

where $w = Z - z$. w is thus the width of the band. Setting the partial derivatives of (52) with respect to z (with w constant), and w (with z constant) equal to zero, yields:

$$\partial E(TC)/\partial z = B_0\ tx^2/z^2w + 2r/3 = 0 \tag{53}$$

$$\partial E(TC)/\partial w = -B_0\ tx^2/zw^2 + R/3 = 0 \tag{54}$$

11 In this case, cash holdings follow a random walk without drift.

which yield the optimal values z^* and w^* as:

$$z^* = (3B_0\ tx^2/4R)^{1/3} \tag{55}$$

$$w^* = 2z^* \tag{56}$$

Since $w = Z - z$, (56) implies that the optimal upper limit Z^* will be:

$$Z^* = 3z^* \tag{57}$$

Equation (57) specifies the *relative* width of the band between the upper and lower limits as:

$$(Z - z)/z = 2$$

which is independent of the interest rate and the brokerage cost, though, by (55) to (57), the *absolute* width of the band does depend upon these variables.

From (57), the mean buffer stock balances M^{b} under the assumptions of this model are given by:

$$M^{\mathrm{b}} = (Z + z)/3 \tag{58}$$

Therefore, the average optimal buffer stock balances $M^{\mathrm{b}*}$ derived from (55), (57) and (58) are:

$$M^{\mathrm{b}*} = \frac{4}{3}\left(\frac{3B_0\,tx^2}{4R}\right)^{1/3} = \frac{4}{3}\left(\frac{3B_0\sigma^2}{4R}\right)^{1/3} \tag{59}$$

since $\sigma^2 = tx^2$. In (59), the average demand for money depends upon the interest rate and the brokerage cost, as in the certainty version of the transactions demand analysis, and upon the variance of net payments, as in the precautionary demand analysis. The elasticity of the average demand for money with respect to the variance of income is 1/3, and with respect to interest rates it is $(-1/3)$, as in Whalen's (1966) analysis.

However, M–O pointed out that, since there does not exist a precise relationship between σ^2 and Y, there will not exist a precise income elasticity of M^{b} with respect to Y. To illustrate, if we are dealing with a firm's demand for money and Y is its income from sales, a proportionate increase in this sales income due to a proportionate increase in all receipts and payments by it, with their frequency unchanged, will increase x proportionately, so that $\sigma^2 = tY^2$, implying from (59) an income elasticity of 2/3, as in Whalen. But if the amount of each receipt and payment does not change but their frequency is increased, so that t increases proportionately with Y such that $t = \alpha Y$, we have $\sigma^2 = \alpha Yx^2$, thereby implying from (59) an income elasticity of 1/3, again as in Whalen. The implied range for the income elasticity of the average buffer stock balances becomes even larger than from 1/3 to 2/3 if the amounts of the transactions increase while their frequency decreases.

The M–O model extends the analysis of the precautionary demand for money to the case where there are fluctuations within upper and lower limits, with these limits derived in an optimizing framework. The existence of a range for the income elasticity of the average buffer stock balances, rather than a single value as for Baumol's transactions balances, is another empirically appealing feature of the M–O model. These authors considered their model to be especially appealing in explaining the firms' demand for money.

6.6 Buffer stock smoothing or objective models

6.6.1 The smoothing model of Cuthbertson and Taylor

The basic partial adjustment model (PAM), in the context of a single period, often assumes that the adjustment in money balances involves two kinds of costs. One of these is the cost of deviations of actual balances from their desired amount. The second element is the cost of changing the current level of balances from their amount in the preceding period. The one-period first-order PAM[12] assumes that the cost function is quadratic in its two elements, as in:

$$TC = a(M_t - M_t^*)^2 + b(M_t - M_{t-1})^2 \tag{60}$$

where:

TC = present discounted value of the total cost of adjusting balances
M = actual money balances
M^* = desired money balances.

The buffer stock models (Carr and Darby, 1981; Cuthbertson, 1985; Cuthbertson and Taylor, 1987; and others) posit an *intertemporal* cost function rather than a one-period one and minimize the present expected value of this cost over the present and future periods. This implies taking account of both types of costs over the present as well as the future periods. Hence, for the first element of cost, the expected cost of future deviations of actual from desired balances, in addition to the cost of a current deviation, is taken into account. This modification allows the agent to take account of the future levels of desired balances in determining the present amounts held. For the second element of cost, the justification for the intertemporal extension is as follows: just as last period's money holdings affect the cost of adjusting this period's money balances, so would this period's balances affect the cost of adjusting next period's balances, and so on, and these future costs attached to current money balances need to be taken into consideration in the current period. The resulting cost function is intertemporal and forward looking.

The intertemporal extension of (60) for $i = 0, 1, \ldots, T$ is:

$$TC = \Sigma_i D^i \left[a(M_{t+i} - M_{t+i}^*)^2 + b(M_{t+i} - M_{t+i-1})^2 \right] \tag{61}$$

where:

D = gross discount rate $(= 1/(1+r))$.

In equation (61), a is the cost of actual balances being different from desired balances and b is the cost of adjusting balances between periods. b can be the brokerage cost of selling bonds but, as we have seen in earlier discussions, this can be more than just a monetary cost. The economic agent is assumed to minimize TC with respect to $M_{t+i}, i = 0, 1, \ldots, T$. Its Euler condition for the last period T is:

$$\partial TC/\partial M_{t+T} = 2a(M_{t+T} - M_{t+T}^*) + 2b(M_{t+T} - M_{t+T-1}) = 0$$

12 See Chapter 7 for the treatment of one-period partial adjustment models in the context of money demand functions.

so that:

$$\begin{aligned} M_{t+T} &= \frac{a}{a+b} M^*_{t+T} + \frac{b}{a+b} M_{t+T-1} \\ &= A_1 M^*_{t+T} + B_1 M_{t+T-1} \end{aligned} \tag{62}$$

Where $A_1/(a+b)$ and $A_1 + B_1 = 1$. For $i < T$, the first-order cost-minimizing conditions are:

$$\frac{\partial C}{\partial M_{t+i}} = 2a(M_{t+i} - M^*_{t+i}) + 2b(M_{t+i} - M_{t+i-1}) - 2b(M_{t+i+1} - M_{t+i}) = 0$$

$$\begin{aligned} M_{t+i} &= \frac{a}{a+2b} M^*_{t+i} + \frac{b}{a+2b} M_{t+i-1} + \frac{b}{a+2b} M_{t+i+1} \\ &= A_2 M^*_{t+i} + B_2 M_{t+i-1} + B_2 M_{t+i+1} \end{aligned} \tag{63}$$

where $A_2 + 2B_2 = 1$. In (63), both the future and past values of actual balances M, as well as the future values of M^*, affect the demand for money in each period. (63) implies[13] that:

$$M_t = q_1 M_{t+i-1} + (a/b) q_1 \quad \Sigma_i q_i^i M^*_{t+i} \tag{64}$$

where $q_1 + q_2 = (a/b) + 2$ and $q_1 q_2 = 1$. We need to specify the demand function for the desired money balances M^*. As derived in Chapters 2 and 3, assume it to be:

$$M^*_{t+i}/p_{t+i} = b_y y_{t+i} + b_R R_{t+i} \tag{65}$$

Further, in the context of uncertainty and using the expectations operator E_{t-1} for expectations held in $t-1$, let:

$$M_t = E_{t-1} M_t + M_t^u + \mu_t \tag{66}$$

13 For the derivation, see Cuthbertson (1985, pp. 137–8; Cuthbertson and Taylor, 1987, pp. 187–8). The following derivation is from Cuthbertson (1985, pp. 137–8). The steps are:
Multiply (63) by $(a+2b)$ and rearrange to:

$$(a + 2b - bL - bL^{-1}) M = aM^*$$

where L^{-1} is the forward (as opposed to the lag) operator, so that $L^{-1} M_t = M_{t-1}$. Multiply by L and divide both sides by b. This gives:

$$\left[-L\left(\frac{a+2b}{b}\right) + L^2 + 1 \right] M = -\frac{a}{b} M^*_{-1}$$

$$\left[-L\left(\frac{a+2b}{b}\right) + L^2 + 1 \right] = (1 - q_1 L)(1 - q_2 L)$$

$$= 1 - (q_1 + q_2) L + q_1 q_2 L^2$$

where $q_1 + q_2 = (a/b) + 2$ and $q_1 q_2 = 1$. Assume $q_2 > 1$, so that $q_1 < 1$. Hence,

$$(1 - q_1 L) M = (1a/b)(1 - q_2 L)^{-1} M^*_{-1}$$

$$= (-a/b)(1 - q_1^{-1} L)^{-1} M^*_{-1}$$

Using the Taylor expansion, $(1 - \lambda L)^{-1} = -(\lambda L)^{-1} - (\lambda L)^{-2} - \cdots$, we get:

$M_t = q_1 M_{t-i-1} + (a/b) q_1 \quad \Sigma_i \, q_i^i M^*_{t+i}$

where M_t^{u} has been introduced to take account of errors in the expected value of M_{t+i}^* due to unexpected changes in its determinants in (65). From (64) to (66),

$$M_t = q_1 M_{t-1} + (a/b) q_1 \Sigma_i \, q_i^i \{ b_y y_{t+i} + b_R R_{t+i} \} p_t + M_t^{\mathrm{u}} + \mu_t \tag{67}$$

In (67), the actual demand for money depends upon the future and current values of income and interest rates, which shows the model to be a forward-looking one. It also depends upon the lagged value of money balances, thus incorporating a backward-looking element. The model is, therefore, both forward *and* backward looking. Note that the estimation of (67) will require the prior specification of the mechanism for deriving the expected future values of y and R, and also of the mechanism for the estimation of M^{u}. The estimation procedures for these are discussed in Chapter 7.

6.6.2 The Kanniainen and Tarkka (1986) smoothing model

An alternative version of the intertemporal adjustment cost function (61), used by Kanniainen and Tarkka (K–T) (1986),[14] is:

$$TC = E_t \Sigma_i \left[D^i \left\{ a(M_{t+i} - M_{t+i}^*)^2 + b(z_{t+i})^2 \right\} \right] \quad i = 0, 1, 2, \ldots \tag{68}$$

where z_t are the "induced" changes in money balances brought about by the agent's own actions and b is interpreted as the brokerage cost of converting bonds to money. The other variables are as defined earlier, with M being nominal balances, M^* the steady-state desired balances and D the discount factor. The rationale for this specification of the cost function is that while the induced changes in money holdings impose a brokerage cost, the autonomous changes do not since they result from the actions of others. The adjustment in nominal balances in t from those in $t-1$ occurs due to autonomous and induced changes in t, so that:

$$M_t - M_{t-1} = z_t + x_t \tag{69}$$

where:

z_t = induced changes in money holdings
x_t = autonomous changes.

Substitute (69) in (68) and, to minimize total cost, set the first-order partial derivatives of the resulting equation with respect to m_{t+i}, $i = 0, 1, 2, \ldots$, equal to zero. This process yields the Euler equations as:

$$E_t M_{t+i+1} - \beta E_t M_{t+i} + (1+R) E_t M_{t+i-1} = -\alpha E_t M_{t+i}^* + E_t x_{t+i+1} - (1+R) E_t x_{t+i} \tag{70}$$

where:

$\alpha = a/Db$
$\beta = (1/D)\{(a/b) + D + 1\}$
$i = 0, 1, 2, \ldots.$

Note that (70) represents a large number of equations and shows the extensive information requirements of such models. To determine money demand in period t, the agent must have

14 The following exposition is based on Kanniainen and Tarkka (1986) and Mizen (1994, pp. 50–51).

expectations on the autonomous changes in money holdings in $(t+1)$ and its optimal money balances, with the latter requiring this information for period $(t+2)$, and so on. With new information becoming available each period, the model will require continual recalculation.

Equation (70) is a stochastic second-order difference equation in $E_t\ M_{t+i}$. Its roots are:

$$\lambda_1, \lambda_2 = (1/2)\left[\beta \pm \left\{\beta^2 - 4(1+R)\right\}^{1/2}\right]$$

with $\lambda_1 > 0$ being the stable root and $\lambda_2 < 0$ being the unstable one. The latter was ignored by K–T in order to exclude cyclical adjustment. Using the positive root λ_1, the Euler condition becomes:

$$E_t M_{t+1} = \lambda_1 M_t + \left[\lambda_1 \alpha/(1+R-\lambda_1)\right] M_t^* - \Sigma_i \left[\lambda_1/(1+R)\right]^i \left[E_t x_{t+i-1} - (1+R) E_t x_{t+i}\right] \tag{71}$$

In (71), the impact of the autonomous adjustment x_t on money demand is given by λ_1, the stable root of (70). This impact is the same whether it was anticipated or not.[15] The impact of future autonomous shocks, on which expectations have to be formed, depends upon the rate of time preference. If this rate is high, these expectations will have to be formed for only some periods ahead. Further, changes in these expectations will shift the money demand function.

Substitute (71) into (70) and solve for M_t, noting that $E_t M_t^* = M_t^*$. This yields:

$$M_t = \lambda_1 M_{t-1} + \rho M_t^* + \lambda_1 x_t + z_t \tag{72}$$

where $\rho = [\lambda_1 (a/b)(1+R)/(1+R-\lambda_1)]$ and z_t^* represents the weighted sum of the future shocks to net receipts and payments. z_t^* is given by:

$$z_t^* = -(1-\lambda_1)\, \Sigma_i \left\{\lambda_1/(1+R)\right\}^i E_t x_{t+j} \tag{73}$$

The above model can be transformed into real terms by dividing (71) by the current price level p_t. The resulting equation, based on (71),[16] is:

$$\ln m_t = a_0 + (1-\lambda_1) \ln m_t^* + \lambda_1 \ln m_{t-1} + \gamma \ln p_t/p_{t-1} + \lambda_1 x_t/M_{t-1} + z_t/M_{t-1} \tag{74}$$

where $m_t = M_t/P_t$ and $m_t^* = M_t^*/P_t$.

15 This conclusion differs from that of Carr and Darby (1981) and Santomero and Seater (1981).

16 The procedure is as follows. Subtracting M_{t-1} from both sides of (72) gives:

$$M_t - M_{t-1} = (1-\lambda_1)\left[\rho/(1-\lambda_1) M_t^* - M_{t-1}\right] + \lambda_1 x_t + z_t$$

Divide both sides of this equation by M_{t-1} and use the approximation $\ln(1+n) \cong n$ for small values of n. This gives:

$$\ln M_t = \ln\left[\rho/(1-\lambda_1)\right]^{1-\lambda_1} + (1-\lambda_1) \ln M_t^* + \lambda_1 M_{t-1} + \lambda_1 x_t/M_{t-1} + z_t/M_{t-1}$$

Now subtract $\ln p_t$ from both sides and also add and subtract the term $\lambda_1 \ln p_{t-1}$. This will give (74) with $a_0 = \ln\left[\rho/(1-\lambda_1)\right]^{1-\lambda_1}$. See K–T for the derivations of these equations and those reported in the text.

K–T specified the desired demand m_t^* as a log-linear function of y_t and R_t, such that:

$$m_t^* = \gamma y_t^{\theta} R_t^{\eta} \tag{75}$$

Where θ and η are parameters. The critical autonomous net payments variable x_t was defined as:

$$x_t = \Delta L_t + \Delta L_t^g + B_t \tag{76}$$

where:

L = domestic credit expansion
L^g = government net borrowing from abroad
B = surplus in the balance of payments on current account

On the future values of x_t, the following extrapolative model was assumed:

$$E_t x_{t+i} = x_t(1+\theta)^i$$

where θ can be positive or negative or zero. Assuming z_{t+i} to be proportional to x_{t+i} such that $Z_{t+i} = -\xi X_{t+i}$, K–T use (74) to specify the estimating equation as:

$$\ln m_t = a_0 + (1-\lambda_1)\ln m_t^* + \lambda_1 \ln m_{t-1} + \gamma \ln p_t/p_{t-1} + (\lambda_1 - \xi)\, x_t/M_{t-1} + \mu_t \tag{77}$$

where μ is random noise. Note that the current autonomous injections of money increase current money holdings through the variable x_t.

The differences between (67), which is the estimating equation for Cuthbertson and Taylor, and (77), which is the estimating equation for K–T, arise from their different cost functions. (77) is derived from (68), which assumes that only induced changes in money balances impose adjustment costs, whereas (67) is derived from (61), which attaches such costs to the total difference between balances in periods $t+i$ and $t+i-1$ and, as such, requires a wider notion of adjustment costs than merely brokerage costs. Both models are forward (and backward) looking models and require specification of the procedures for estimating the future values of the relevant variables. They also require specification of the estimation procedure for net payments and receipts. Part of these net payments and receipts can be anticipated and part unanticipated.

The two approaches of Cuthbertson *et al* and of Kanniainen and Tarkka are similar in many ways. Both are examples of smoothing buffer stock models in which money holdings are free to vary from their desired levels. Such models illustrate some of the most common elements used for specifying the buffer stock analysis of the demand for money. Their critical feature is that the autonomous injections of money supply in the economy are, for some time, passively absorbed by the public in actual money holdings.

To give an indication at this stage of the empirical findings on the critical points in the above analysis, we here briefly mention Kanniainen and Tarkka's empirical results. They estimated their model for five industrialized economies (West Germany, Australia, USA, Finland and Sweden) for the period 1960–82. The estimated coefficients of their model had the predicted signs and plausible magnitudes. The lagged money variable had a magnitude consistent with other studies. However, as indicated by their estimate of γ, the adjustment of money balances to changes in prices was found to be costly. Their buffer stock equation performed better than the standard (non-buffer stock) money demand models: the coefficients of the injection

variable x_{t+i} were positive and significant, thus supporting the buffer stock approach. These findings also support the hypothesis that the monetary injections from different sources are first absorbed in nominal money holdings and then dissipated.

6.7 Empirical studies on the precautionary and buffer stock models

While the transactions, speculative and precautionary models determined a unique optimal demand for money for each component, the buffer stock models allowed fluctuations in money holdings either in a band or around an optimal long-run path. These models are forward (and backward) looking and require specification of the procedures for estimating the future values of the relevant variables. They also require specification of the estimation procedure for net payments and receipts. Part of these net payments and receipts can be anticipated and part unanticipated. Another feature of the buffer stock models for estimation purposes is that the unanticipated injections of money supply in the economy are, for some time, passively absorbed by the public.

There are two broad types of empirical studies of the buffer stock money demand. One of these distinguishes between a long-run (planned or permanent) desired money demand and a short-run (buffer stock or transitory) money demand, and estimates their sum by standard regression techniques. The empirical works of Darby (1972), Carr and Darby (1981) and Santomero and Seater (1981), among others, belong in this category. We shall refer to this category as the shock-absorption money-demand models. The second type of empirical studies uses cointegration techniques and error-correction modeling. Since these techniques will be discussed in Chapter 8, which also reports on their findings on money demand, this chapter reports on only the former type of studies.

Shock-absorption money demand models

Darby (1972) proposed and tested a version of the "shock-absorption" model of money demand. In setting up his model, he argued that most of any positive transitory saving will be initially added to money balances and then gradually reallocated to other assets or be depleted by subsequent negative transitory saving – with money balances reverting to their long-run desired ("permanent") levels at the end of these adjustments. Money balances therefore, act to absorb shocks in income and saving. The Darby shock-absorber model is an early version of the buffer stock models of money and limits its shocks to innovations in income.

Darby separated money holdings into two categories, permanent and transitory, as in:

$$M_t = M_t^{\mathrm{P}} + M_t^{\mathrm{T}} \tag{78}$$

The demand for permanent money balances was assumed by Darby to be:

$$M_t^{\mathrm{P}} = \alpha_0 + \alpha_y Y_t^{\mathrm{P}} + \alpha_{RL} RL_t + \alpha_{RS} RS_t + \alpha_{RM} RM_t \tag{79}$$

where:

M^{P} = permanent real balances
M^{T} = transitory real balances
Y^{P} = permanent real income
Y^{T} = transitory real income
RL = long-term nominal interest rates
RS = short-term nominal interest rate
RM = nominal yield on money balances.

Equation (79) specifies the dependence of permanent balances on permanent income and various interest rates.

For transitory balances M^{T}, Darby assumed that:

$$\Delta M_t^{\mathrm{T}} = \beta_1 S_t^{\mathrm{T}} + \beta_2 M_{t-1}^{\mathrm{T}} \quad 0 > \beta_1 > 0, \quad \beta_2 < 0 \tag{80}$$

where the proportion β_1 of transitory real saving S^{T} is added to transitory real balances during the period, but last period's transitory balances are run down or adjusted at the rate β_2.

Darby used Milton Friedman's ideas on permanent and transitory income in which:

$$Y_t = Y_t^{\mathrm{P}} + Y_t^{\mathrm{T}} \tag{81}$$

where permanent and transitory income are not correlated with each other and permanent income is generated by an adaptive expectations procedure. Further,

$$S_t^{\mathrm{T}} = Y_t^{\mathrm{T}} - C_t^{\mathrm{T}} \tag{82}$$

where transitory consumption C^{T} is an independent random variable with a zero mean, so that Y_t^{T} was substituted for S_t^{T} in the estimating equation. As mentioned above, a proportion β_1 of it is accumulated in transitory money balances during the period and eventually reallocated to other assets.

Equations (78) to (82) imply that:

$$M_t = \alpha_0(1-\beta) + \beta_1 Y_t^{\mathrm{T}} + \beta M_{t-1} + \alpha_Y Y_t^{\mathrm{P}^*} + \alpha_{RL} RL_t^* + \alpha_{RS} RS_t^* + \alpha_{RM} RM_t^* \tag{83}$$

where $\beta = (1+\beta_2)$, $Y^{\mathrm{P}^*} = (1-\beta)Y_t^P$, $RL_t^* = (1-\beta)RL_t$, $RS_t^* = (1-\beta)RS_t$ and $RM_t^* = (1-\beta)RM_t$.

Darby's finding for USA for the period 1947:1 to 1966:4 was that β_1 was about 40 percent, so that transitory income and saving had a strong effect on money balances and transitory balances increased by about 40 percent of transitory income. β_2, the induced reduction in transitory balances per quarter, was about 20 percent. These findings support the buffer stock approach where net income receipts are temporarily added to money balances and then gradually adjusted at periodic intervals. The estimated adjustment is relatively slow, though Darby also found that both β_1 and β_2 had increased since the 1940s. With the increasing innovation in the financial markets in recent decades, the increase is likely to have continued and be quite significant.

While the above model introduces the notion of transitory money balances arising from transitory income and saving into the analysis, it does not deal with the differing effects of anticipated and unanticipated changes in the money supply, and therefore does not deal with innovations in money supply. Carr and Darby (1981) argue that the anticipated changes in money supply are integrated by economic agents into their decisions on consumption, etc. and are therefore already incorporated into the current price level, with real balances held being unaffected by the changes in the price level and the anticipated money changes. However, the unanticipated money supply change alters the net receipts of the public and can be treated as an element of transitory income. It may be wholly or partly added to buffer balances, is thereby not spent and is not reflected in the price level. Hence, changes in the unanticipated money supply alter real balances while changes in the anticipated money supply do not.

To incorporate these arguments into the analysis, Carr and Darby (1981) assumed that:

$$M_t^{s} = M_t^{s*} + M_t^{su} \tag{84}$$

where:

M^{s} = nominal money supply
M^{s*} = anticipated nominal money supply
$M^{su} = M^{s} - M^{s*}$ = unanticipated nominal money supply.

The short-run desired demand function was specified in real terms, with a partial adjustment model, as:

$$m_t^{d} - m_{t-1} = \lambda (m_t^{*} - m_{t-1}) \tag{85}$$

where:

m^{d} = short-run desired demand for real balances (log scale)
m^{*} = long-run desired demand for real balances (log scale)

so that the desired short-run demand for real balances is:

$$m_t^{d} = \lambda m_t^{*} + (1 - \lambda) m_{t-1} \tag{86}$$

where the short-run desired demand for real balances is a weighted average of the long-run demand and one-period lagged balances. The actual holdings of money balances are the sum of the short-run desired balances, transitory income and unanticipated money supply. Hence:

$$m_t = \lambda m_t^{*} + (1 - \lambda) m_{t-1} + \beta y_t^{T} + b M_t^{su} \tag{87}$$

The long-run desired demand is given by:

$$m_t^{*} = \gamma_0 + \gamma_1 y_t^{P} + \gamma_2 R_t \tag{88}$$

Therefore:

$$m_t = \lambda\gamma_0 + \lambda\gamma_1 y_t^{P} + \lambda\gamma_2 R_t + (1 - \lambda) m_{t-1} + \beta y_t^{T} + b M_t^{su} \tag{89}$$

Permanent and transitory income were measured as in the Darby model earlier; in calculating permanent income, the weight on the current quarterly income was set at 0.025 percent. In the present model, the demand for real balances depends upon transitory income and unanticipated money supply changes. The theoretical arguments require their coefficients to be positive.

Carr and Darby (1981) tested this model for eight industrial countries (USA, UK, Canada, France, Germany, Italy, Japan and the Netherlands) for the period 1957:1 to 1976:4 and reported that the coefficient b on unanticipated money supply was significant and ranged between 0.7 and 1 for all countries, using generalized least-squares (GLS) estimates. The coefficient β was significant and positive for the USA but was not significant for the other countries. To illustrate, using GLS estimates for the coefficients, the estimated value of β (the coefficient for transitory income) was 0.090 while that of b (the coefficient for unanticipated money supply) was 0.803 for the USA; the corresponding estimate of β for Canada was 0.018, which was not significant, but the estimate of b was 0.922 which was significant. Hence, the influence of transitory income on money balances was much weaker than that

of unanticipated money supply changes, with most of the latter added to money balances in the current quarter, so that the impact effect of such changes on the price level or economic activity would be minimal.

Santomero and Seater (1981) started with the Whalen (1966) model and introduced elements of search theory into their buffer stock model. They assumed that an individual with buffer balances in currency and demand deposits will search for alternative assets in a context of incomplete information on such assets – especially long assets and durable goods which are bought infrequently – but there is a cost to acquiring more information. Given this cost, the individual does not continuously perform the cost minimization decision to buy the alternative assets but only does so at discrete points in time, while holding buffer stocks in the intervals between decisions. The source of the shocks inducing a change in the pattern of assets held is among the determinants of this cost, as are interest rates, past shocks, variance of transactions, etc. Santomero and Seater showed that, under their assumptions, excess money balances are run down gradually rather than completely at each decision point.

The empirical analysis of Santomero and Seater was as follows:

$$M_t = M_t^* + M_t^{\mathrm{T}} \tag{90}$$

where:

M = short-run desired real balances
M^* = long-run (equilibrium) desired real balances
M^{T} = transitory real balances.

M^* was assumed to depend upon permanent income and the cost differential of holding money rather than other assets, in a Cobb–Douglas form, as:

$$M_t^* = \alpha Y_t^{\mathrm{P}\beta}(R_{1t} - R_{Mt})^{\rho} \cdots (R_{mt} - R_{Mt})^{\rho} \tag{91}$$

The general determinants of transitory balances were specified as:

$$M_t^{\mathrm{T}} = M^{\mathrm{T}}(S_t, S_{t-1}, \ldots, (R_A - R_M), \beta_t) \tag{92}$$

where S_t was the shock to real balances in t, R_A was the nominal interest rate on alternative assets (savings deposits), R_M was the nominal interest rate on money and β was Whalen's penalty cost on holding inadequate balances. (92) was given the more specific form:

$$M_t^{\mathrm{T}} = DM_t^* \tag{93}$$

where the disequilibrium factor D was a distributed lag function of past transitory shocks. It was assumed that there were two sources of shocks, one to income and the other to the money supply. The specific form for D was:

$$D = \sum_{j=0}^{\mathrm{N}} Z^j \left[\frac{(Y_{t-j} - Y_{t-j}^{\mathrm{P}}) + (M_t^{\mathrm{s}} - M_{t-1}^{\mathrm{s}})}{Y_{t-j}^{\mathrm{P}}} \right] \tag{94}$$

where:

Y = current real income
Y^{P} = permanent real income
M^{s} = real money supply.

Equation (94) assumes that all shocks, whether from income or money supply changes, have the same pattern of effects on transitory balances. Further, innovations in money demand are not considered, so that either they do not occur[17] or real balances adjust instantly to them. If the money demand function is unstable, (94) should be modified to include shifts in money demand.

The estimated demand function for short-run balances implied by (90) to (94) is of the form:

$$M_t^{\mathrm{d}} = \alpha\, Y_t^{\mathrm{P}\beta}\, (R_{1t} - R_{Mt})^{\gamma} \cdots (R_{mt} - R_{Mt})^{\rho} \times \left[1 + \sum_{j=0}^{\mathrm{N}} Z^j \left(\frac{(Y_{t-j} - Y_{t-j}^{\mathrm{P}}) + (M_t^{\mathrm{s}} - M_{t-1}^{\mathrm{s}})}{Y_{t-j}^{\mathrm{P}}} \right) \right]^{\delta} \tag{95}$$

where $\alpha, \beta, \delta > 0$ and $\gamma, \rho < 0$, and R_i is the nominal rate of return on the ith asset.

The above model was estimated for M1 and M2 for the USA for the period 1952:2 to 1972:4, using the Cochrane–Orcutt technique to eliminate first-order serial correlation. It was assumed in the estimation process that equilibrium was achieved within each quarter between the money supply and the short-run money demand, so that the latter was proxied by the money supply. Only two interest rates – the commercial paper rate and the commercial passbook rate – were used. The estimate of the coefficient Z was significant and positive for both M1 and M2. Hence, both transitory income and changes in the money supply had a short-run positive impact on short-run money demand, thereby showing evidence of buffer holdings of money balances. Further, transitory balances did not increase proportionately with the magnitude of the shock, so that large shocks were corrected faster than smaller ones. M1 and M2 holdings adjusted within two to three quarters to their desired levels, implying a fairly fast rate of adjustment.

A microfoundations search model of precautionary balances

Faig and Jerez (2007) use microeconomic optimizing foundations to model the demand for money as a demand for precautionary balances held against uncertain expenditure needs. They build a search model incorporating shocks to individuals' preferences. Individuals decide on their money balances prior to the unknown preference shock, with large preference shocks assumed to be rarer than small ones. At low interest rates, individuals hold enough balances so that their consumption purchases are not very liquidity constrained, but they are willing to allow their purchases to be liquidity constrained to a greater extent at higher interest rates, so that velocity falls at higher interest rates. Their empirical estimates for the USA for the periods 1892 to 2004 capture the fall in the velocity of M1 during the low-interest period of the Great Depression and its rise from the mid-1960s to the mid-1980s, which had high interest rates. Financial innovations, such as credit cards, Internet banking, etc., as well as the reduction in brokerage costs (which has reduced the cost of transfers between M1 and other assets), have meant that households can better accommodate unexpected expenditures, even with inadequate money balances. These developments have

17 Santomero and Seater (1981) assume that the money demand function is stable. In fact, empirical studies in recent decades have shown it to be unstable and therefore to generate transitory excess money holdings.

reduced the need for precautionary balances. Therefore, the empirical impact of financial innovations and the IT revolution has been to reduce the demand for M1 and increase its velocity.

Conclusions

This chapter has presented some of the basic models of the precautionary demand for money. This demand arises because of the uncertainty of the timing of payments and receipts, so that a major determinant of such demand is the probability distribution of net payments. Whalen's analysis captured this by the variance of the distribution, under the assumptions that the distribution was normal and the individual wanted to keep balances equal to a specified proportion of the variance of the distribution. How much of this proportion is kept by a particular individual will depend upon his degree of risk aversion. This analysis shows that the elasticities of precautionary demand with respect to interest rates and income are not 1/2, as were the elasticities for transactions demand under certainty of the timing of receipts and payments, but are likely to be lower.

The precautionary demand analysis also brings the interest rates on stand-by credit facilities such as overdrafts and trade credit, and the penalties for being caught short of a payment medium, into the determinants of the precautionary demand and through it into those of the total demand for money. Such facilities and penalties differ between households and firms, and between large and small firms. Further, they often also differ by industry, so that we should expect the demand functions for money to differ between sectors and industries.

Is the precautionary demand for money a significant part of the total money balances? The answer to this will depend upon the degree of uncertainty of income and expenditures, and the relative penalty costs of being caught short of funds. Some numerical calculations done by Sprenkle and Miller suggest that the precautionary balances can be about one-third or more of the transactions balances. Further, with the increasing availability of short-term assets, which offer a higher return without a significantly higher risk than M1, the speculative demand for M1 would nowadays be insignificant, so that the precautionary demand could be greater than the speculative demand. Consequently, the study of the precautionary demand for money has become more prominent in recent years. On the negative side, the increasing availability of several close substitutes for M1 means that the precautionary needs of the individual could also be met by the holdings of such near-monies, so that the precautionary demand for M1 would also be small.

Since the precautionary demand for money reflects the influence of the uncertainty of incomes and expenditures, fluctuations in this degree of uncertainty over the business cycle would imply fluctuations in the precautionary balances over the cycle. Boom periods of high employment and low uncertainty of income would mean a lower precautionary demand for money, for given income levels, than recessionary periods with greater uncertainty of keeping one's job. A higher variance of the rate of inflation would imply a higher variance of net receipts and therefore a higher precautionary demand for money. Provision of national health care (medicare) under which no payment has to be rendered for medical services reduces both precautionary savings and precautionary money demand.

The buffer stock approach constitutes a very significant innovation in money demand modeling and represents an extension of the notions behind the precautionary demand for money. However, this approach goes further than the pure precautionary demand motive by recognizing that there are different costs of adjusting various types of flows and stocks, and that, for the individual, adjustment in money balances is often the least cost immediate option

for many types of shocks. The result is that money balances are increased and decreased as a buffer in response to many types of shocks, and are only adjusted to their long-run equilibrium levels as and when it becomes profitable to adjust other stocks and flows. Hence, a distinction has to be drawn not only between short-run desired money balances and long-run ones, but also between the former and the balances actually being held. The difference between these concepts is buffer stock balances. Actual balances will be larger than short-run desired balances for positive (unanticipated) shocks to the money supply and to income, and smaller for negative shocks in the latter. Hence, unlike the standard money-demand models, the buffer-stock models imply the dependence of money demand on the shocks to money supply, so that short-run money demand is not independent of money-supply shifts.

The divergence of actual balances from the short-run desired balances and of the latter from the long-run balances provides one explanation for the delayed response of nominal income and interest rates to monetary policy where the latter include unanticipated changes in money supply. Hence, in terms of the implications of the buffer-stock models for monetary policy, these models imply that since part of the changes in the money supply are accommodated through passive money holdings, the impact of such changes on market interest rates is correspondingly smaller and the full impact takes some time to occur. Correspondingly, the full impact of such money supply changes on nominal national income takes several periods and is larger than the short-term effect.

While the rule models, with money balances fluctuating in a band with upper and lower limits, inaugurated the buffer stock notion through the pioneering contribution of Miller and Orr (1966), empirical work has tended to follow the ideas of the smoothing models.

The empirical work of Carr and Darby (1981) provides comparison between the relative effects of income shocks and money-supply shocks. The effect of the income shocks on money demand is weaker and, for many countries, insignificant while the effect of the money-supply shocks is significant and substantially stronger. The latter represents a confirmation of the buffer stock hypothesis.

These findings imply that – since some part of the changes in the money supply result in passive money holdings, which are then gradually eliminated over time – the impact of such money supply changes on market interest rates is correspondingly smaller and the full impact takes some time to occur. This has been confirmed in the findings, reported in Chapter 9, from many error-correction models. Further, the full impact of money supply changes on nominal national income will also take several quarters and the overall effect will be larger than the short-term one.

Summary of critical conclusions

- Precautionary savings and precautionary money demand arise because of the uncertainty of either income or expenditures.
- Precautionary money demand depends on the variances of income and expenditures and the availability of overdraft facilities, as well as on the penalty cost of a shortfall in money holdings.
- Buffer stock money demand arises because money has a lower cost of adjustment than commodities, labor and leisure. Money acts as a passive short-term inventory of purchasing power until it is optimal to make adjustments in the latter set of variables.
- Rule models of buffer stock money demand allow fluctuations in money demand in a range between pre-selected upper and lower limits.

- ❖ Smoothing models of buffer stock money demand imply that actual money holdings will vary around their desired long-run level.
- ❖ Precautionary and buffer stock holdings of money would vary with the phase, duration and amplitude of the business cycle, and with unemployment rate.
- ❖ Precautionary and buffer stock holdings of money also depend on the availability of other highly liquid interest-bearing assets in the economy. In the presence of these, such holdings could be insignificant.
- ❖ Empirical evidence confirms the existence of buffer stock holdings of money balances. Therefore, money demand is not independent of money supply.

Review and discussion questions

1. "Individuals hold money because of uncertainty over the timing of transactions. Therefore, the theory of the transactions demand for money must take account of this uncertainty." Discuss this statement.

   How can this uncertainty be incorporated into a model utilizing the inventory analysis of the transactions demand for money? Present a model that does so.
2. What is the buffer stock demand for money and how does it differ from the precautionary money demand? Present at least one model of each that shows such a difference.
3. If the speculative demand for money is zero in a financially developed economy, as some have claimed, is the precautionary demand for money also zero? Evaluate with reference to both M1, M2 and broader monetary aggregates. If both the speculative and precautionary components of money demand are zero, what about the transactions demand? Consider both households and firms in your answer.
4. "Some recent empirical studies seem to show that the money demand function may not be independent of the money supply function." Report on the methodology and results of at least one such study.
5. If money demand is dependent on changes in the money supply in the short-run, as the buffer stock models show, does the functional form of the money demand function remain the same or change? What are the arguments of the money demand function incorporating a buffer stock component? How would you estimate this function?
6. What is the justification for and what are the arguments against the buffer stock approaches to money demand?
7. To what extent is Baumol's inventory-theoretic approach, with its assumption of certainty, a satisfactory explanation of money demand?

   How does managing an inventory of money differ from managing inventories of goods? How do such differences affect the speed of adjustment of money demand to expected and unexpected changes in money supply?
8. Present at least one rule model and one smoothing model of the buffer stock demand for money. What are the major differences in their implied money demand functions?

References

Akerlof, G.A., and Milbourne, R.D. "The short-run demand for money." *Economic Journal*, 90, 1980, pp. 885–900.

Carr, J., and Darby, M.R. "The role of money supply shocks in the short-run demand for money." *Journal of Monetary Economics*, 8, 1981, pp. 183–99.

Cuthbertson, K. *The Supply and Demand for Money*. Oxford: Basil Blackwell, 1985, pp. 30–2, 35–9, 130–43.

Cuthbertson, K., and Taylor, M.P. "The demand for money: a dynamic rational expectations model." *Economic Journal*, 97, 1987, pp. 65–76.

Darby, M.R. "The allocation of transitory income among consumers' assets." *American Economic Review*, 62, 1972, pp. 928–41.

Faig, M., and Jerez, B. "Precautionary balances and the velocity of circulation of money." *Journal of Money, Credit and Banking*, 39, 2007, pp. 843–73.

Kanniainen, V., and Tarkka, J. "On the shock-absorption view of money: international evidence from the 1960's and 1970's." *Applied Economics*, 18, 1986, pp. 1085–101.

Milbourne, R. "Disequilibrium buffer stock models: a survey." *Journal of Economic Surveys*, 2, 1988, pp. 187–208.

Miller, M.H., and Orr, D. "A model of the demand for money by firms." *Quarterly Journal of Economics*, 80, 1966, pp. 413–35.

Mizen, P. *Buffer Stock Models and the Demand for Money*. London: St Martin's Press, 1994, Chs. 1–3.

Santomero, A.M., and Seater, J.J. "Partial adjustment in the demand for money: theory and empirics." *American Economic Review*, 71, 1981, pp. 566–78.

Sprenkle, C.M., and Miller, M.H. "The precautionary demand for narrow and broad money." *Economica*, 47, 1980, pp. 407–21.

Whalen, E.L. "A rationalization of the precautionary demand for cash." *Quarterly Journal of Economics*, 80, 1966, pp. 314–24.

7 Monetary aggregation

One of the most persistent questions in monetary economics has been about the proper definition of money. In the nineteenth century, the disputes were about whether demand deposits, gradually increasing in usage, should be included in the definition of money. By the 1950s, their inclusion in money was beyond dispute but new questions had arisen about whether savings deposits should also be in the money measure. While savings deposits are now part of some of the commonly used definitions of money, a fresh set of questions has arisen about the inclusion of other financial assets in monetary aggregates. This perpetual problem with the definition of money and the solutions proposed for it is the subject of this chapter.

Key concepts introduced in this chapter

- Monetary aggregates
- Simple sum aggregation
- Friedman's criterion for defining money
- Weak separability
- Elasticity of substitution among monetary assets
- Variable elasticity of substitution function
- Divisia aggregation
- Certainty equivalence aggregation
- User cost of monetary assets
- Statistical causality
- St Louis monetarist equation

The preceding chapters have discussed several definitions of money. Of these, the *narrow definition* (M1) is currency in the hands of the public plus demand deposits of the public in commercial banks. The broader *Friedman definition of money* (M2) consists of M1 plus time and savings deposits of the public in commercial banks. The still wider definition (M3) includes M2 and adds deposits in near-banks.[1] There are also several variants of M1, M2 and M3, as well as still wider definitions of money, which extend the range of assets that are included to encompass increasingly less liquid assets held by the public. Examples of

1 Beyond M1 and M2, there is no uniformity in defining M3 and M4 and higher Ms, though the increase in digits does indicate the inclusion of progressively less liquid assets in the aggregate.

such assets are Treasury bills and money market mutual funds. The very broad monetary aggregates merge the concept of money in the diffuse concept of "*liquidity.*"

Given the possibility of many definitions of money, the basis on which the definitions are arrived at and their relative validity and performance become essential to any empirical study on money. This basis can be purely theoretical, using the functions of money and concentrating on its role as a medium of exchange and payments. However, as discussed in earlier chapters, this procedure does not normally provide a unique definition of money. Financial assets are created and numerous close substitutes to currency and demand deposits can be easily created in an unregulated, free-enterprise financial system. A plethora of such assets usually exists in developed economies with unregulated financial markets, so that an empirical basis for including some of these in the definition of money and excluding others is needed. One of these procedures is provided by the *theory of aggregation* or composite goods, since any measure of money is an aggregate or composite of its component assets. Aggregation theory requires *weak separability* among the assets to be included in the monetary aggregate, so that a test for weak separability provides a mechanism for judging the validity of the assets to be included in the monetary aggregate.

Once the assets to be included in the definition of money have been selected, the form of the aggregator function over the component assets has to be determined. This function can be specified on an a priori basis or determined by the data. Its common forms in the monetary literature are the *simple sum aggregates,* the *variable or constant elasticity of substitution function* and the *Divisia aggregator function*. The various aggregates in turn have to be tested for their empirical usefulness. The above forms of aggregation and commonly used tests for choosing among them are presented in this chapter.

Section 7.1 points out the failure of monetary theory to provide a unique definition of money when there are several close but not perfect substitutes for currency and demand deposits, and therefore the rest of this chapter examines the empirical considerations that have been proposed for selecting among various empirical measures of money. Section 7.2 discusses Milton Friedman's criterion for empirically defining money. A more rigorous criterion for defining a composite variable such as money is provided by aggregation theory, whose weak separability criterion of aggregation is specified in Section 7.3. The four competing modes of aggregation – simple sum aggregation, variable elasticity of substitution function, Divisia aggregates and certainty equivalence aggregates – are discussed in Sections 7.4, 7.5 and 7.8. A moot question arises about the appropriate cost of holding and using money. If money is used for its liquidity services in financing transactions, the appropriate measure is the user cost of money. This is defined in Section 7.6. Section 7.9 discusses the various criteria for judging among monetary aggregates. Since the Divisia aggregates are among the newer measures and have many appealing features, the appendix to this chapter presents their derivation.

7.1 The appropriate definition of money: theoretical considerations

There are several possible ways of selecting a preferred version among the possible variants of money. One of these ways is the intuitive one of focusing on the functions of money and asserting that the medium-of-payments function is its pre-eminent characteristic, so that only assets that perform this function are entitled to be in the definition of money.[2] In the United

2 The other functions of money are: store of value, standard of deferred payments and unit of account.

States and Canada, until about the 1970s, such a rule would have resulted in the definition of money as currency plus demand deposits (M1) since other assets did not directly perform this role to a significant degree at that time.[3] However, financial developments in the 1970s and 1980s led to the creation of various types of savings deposits on which checks can be written and from which funds can be easily withdrawn, as well as bills paid through automatic teller machines. For many such accounts, further developments in the 1990s allowed transfers to demand deposit accounts or to third parties by telephone or online. Therefore, over recent decades in developed economies, savings and several other types of accounts have come to perform the medium-of-payments function to a greater or lesser extent. Hence the focus on the medium of payments to define money would now support the use of variants of M2 and M3, which not only include savings deposits besides M1 but often also include some other types of liabilities of financial intermediaries.[4]

Therefore, the a priori theoretical specification of the definition of money does not provide a unique measure of money, and economists are forced to look for empirical measures of money. One of the earliest ones, proposed by Milton Friedman and his associates in the 1950s, is presented in the next section.

7.2 Money as the explanatory variable for nominal national income

One of the empirical approaches to the definition of money has been concerned with the policy question: what is the monetary aggregate that can best explain or predict the relevant macroeconomic variables? In several studies in the 1950s and 1960s, Milton Friedman and his associates considered the relevant macroeconomic variable to be nominal national income or expenditures. They argued that the appropriate monetary aggregate is that which is more "closely related" to nominal income than other such aggregates.

This relationship was usually examined by linear or log-linear regressions of the form:

$$Y_t = \alpha_0 + \alpha_1 M_t + \mu_t \tag{1}$$

where Y is nominal national income, M is a monetary aggregate and μ is the disturbance term. The "best" predictor of Y was specified as that monetary aggregate that yields the highest R^2 and also possesses stability of the estimated coefficients. Under this criterion, Friedman and many other economists, in line with the quantity theory, took the relevant macroeconomic variable to be nominal national income. Their empirical findings for the 1950s and 1960s data showed that the "proper" definition of money in many financially developed economies such as the USA, Canada and Britain was currency plus all savings deposits in commercial banks – that is, M2 rather than M1.

Keynesian theorists of the time, and those emphasizing the asset approach, often broadened the list of relevant variables to include an interest rate or rates in addition to nominal

3 In fact, at some of the earlier stages in financial development, while the banks are non-existent or the costs of using demand deposits are very significant, demand deposits themselves do not serve as a medium of payments. In such a stage, the medium of payments function would imply that only currency is money.

4 The history of monetary development is, in a sense, that of the extension of the list of assets that function as the medium of payments. At the early stages, only currency performs this role; then currency and demand deposits; followed by currency, demand deposits and savings deposits; with still more additions of other assets further in the development process.

national income. One application of this was to estimate a linear or log-linear equation of the form:

$$M_t/Y_t = a_0 + a_1 R_t + \mu_t \tag{2}$$

where R is the nominal interest rate. The definition of money arrived at through estimations of (2) usually differed from the Friedman criterion specified by (1), since the two criteria are different.

Further, even if nominal income is the only relevant variable to be explained, financial deregulation and technological innovations since the 1960s have led to shifts in the monetary aggregate that "best" explains nominal national income, with the results depending upon the country as well as the time period of the study. Consequently, there is no clear-cut unique measure of money that has consistently proved to be the "best" one over time for any country or across countries. For instance, during the 1950s and 1960s, as Friedman showed, most studies indicated that, for the USA and Canada, M2 was "more closely" related to national income than M1 in terms of R^2 and the stability of the estimated relationships. But for the 1980s and later decades, M1 appeared to perform better than M2 under this criterion.

7.3 Weak separability

From a rigorous theoretical standpoint, a monetary aggregate is a composite good that must satisfy the following weak separability condition required for aggregation.

Assume that there are n goods, $X_1, \ldots, X_m, X_{m+1}, \ldots, X_n$, whose quantities are related by the function $F(.)$ to X, as in:

$$x = F(x_1, \ldots, x_m, x_{m+1}, \ldots, x_n) \tag{3}$$

where:

x = utility (or output)
x_i = quantity or real value of the ith good
$F(.)$ = utility (or production) function.

Equation (3) can be written as:

$$x = F(f(x_1, \ldots, x_m), x_{m+1}, \ldots, x_n) \tag{4}$$

if and only if F_i/F_j, $i, j = 1, \ldots, m$, is independent of $x_{m+1}, \ldots, x_n$. $F_i = \partial F/\partial x_i$ and F_i/F_j is the marginal rate of substitution (MRS) between X_i and X_j. This independence of the MRS between each pair of goods in a group from all other goods not in the group is known as the *weak separability* of the group from other goods in the overall function. If a collection of goods, $X_1, \ldots, X_m$, satisfies this condition, we can create a composite good M such that its quantity index m can be constructed from the quantities of the goods only in the group, so that we would have $m = f(x_1, \ldots, x_m)$. Further, changes in the quantities or prices of the goods not in the group will not change the relative composition of the composite good. However, they can change the total expenditure on the group and hence the budget constraint for the group.[5]

5 Hence, changes in the price of goods not in the group can cause income effects but not substitution effects in the demand for goods in the group.

Hence, if weak separability holds for $X_1, \ldots, X_m$ from other goods, (4) can be rewritten as:

$$x = F(m, x_{m+1}, \ldots, x_n) \tag{5}$$

where:

$$m = f(x_1, \ldots, x_m) \tag{6}$$

$f(.)$ in (6) specifies a sub-utility function, whose form must be used to construct the quantity (real value) of the aggregate m. To derive the optimal values for $x_1, \ldots, x_m$, the sub-utility function $f(.)$ can be maximized subject to the total expenditures allocated for m, but without reference to the prices or quantities bought of $x_{m+1}, \ldots, x_n$.

Weak separability and monetary aggregation

If the monetary assets $X_1, \ldots, X_m$ are currency, demand deposits and near-money assets, and weak separability from other goods holds for them, they can be grouped into a valid composite monetary aggregate whose quantity depends only on the quantities of its own component assets. The relative demand for the component assets in m would depend only on the quantity index m and the prices of $X_1, \ldots, X_m$, but not directly upon the quantities or prices of the other goods $X_{m+1}, \ldots, X_n$. If $F(.)$ is the individual's utility function over all goods, the remaining goods $X_{m+1}, \ldots, X_n$ could include among them other financial assets, consumption goods, leisure, and any other goods in the individual's utility function. In the case where $F(.)$ is the firm's production function and x is the firm's output, $X_{m+1}, \ldots, X_n$ could include among them the firm's other inputs such as labor, capital, other financial assets and so on.

The condition of weak separability may be satisfied by different groupings of goods, with some of these encompassing others. In such a case, the former would be a wider aggregate and would include the latter – narrower aggregates – as subcategories of composite goods, just as M2 includes M1 as a component.

Note that if the monetary aggregate M, including the assets $X_1, \ldots, X_m$, satisfies the weak separability condition, its functional form will be specified by the *particular* form of $f(x_1, \ldots, x_m)$ that satisfies the weak separability condition. A misspecification in the form of the $f(.)$ function would mean that the weak separability condition is violated in the context of the misspecified function. However, the form of $f(.)$ is normally not known a priori, so empirical work has used a variety of arbitrarily specified forms such as the simple sum aggregate, the variable elasticity of substitution function, the Cobb–Douglas function and more flexible functional forms.[6] Ideally, the functional form applied to the data should be as flexible as possible, so that the data could determine the specific functional form.[7]

6 Among these are the translog form, the Gorman polar flexible functional form and the Fourier flexible functional form.

7 Weak separability is determined directly from the data on quantities (Varian, 1983). Another form of grouping can be based on the quasi-separability of prices. A set of variables is said to be *quasi-separable* if the expenditure on it is a function only of total utility and the prices of goods in the group, but not of the prices of other goods.

Empirical evidence on weak separability of monetary variables

Varian (1983) contributed a pioneering study on judging weak separability among goods, using non-parametric econometric methods. Swofford and Whitney (1987) used Varian's technique for US data (from 1970 to 1985) on consumption goods, leisure and various monetary aggregates. Several definitions of money were found to be weakly separable from consumption goods and leisure, with the broadest measure to do so including currency, demand deposits, checkable deposits and small savings deposits. However, measures broad enough to include money market mutual funds were not weakly separable from consumption and leisure. Conversely, consumption goods and leisure together were weakly separable from monetary assets, but consumption goods alone were not. Our concern is only with the former result. It showed that M1 and M2 were acceptable monetary aggregates but broader measures than M2 were not. Hence, for the period of their study, the definitions or indices of money that include monetary assets beyond those in M2 – and, therefore, implicitly assume their weak separability from other goods in the economy – are misspecified.

Note that the assets that meet the weak separability criterion for inclusion in the monetary aggregate are likely to differ among countries and periods. Further, in view of the considerable degree of innovation and change in the moneyness of assets in the past several decades, the admissible assets in the monetary aggregate have been changing and are likely to have increased beyond M2 for many countries.

Another study using weak separability is Belongia and Chalfant (1989). These authors started with the assumption that monetary assets were weakly separable from consumption and only examined weak separability within this category. Using US monthly data over the brief period from January 1983 to February 1986, they reported that several groups of assets were weakly separable from others. Among these groups were: (C, DD), (C, DD, NOWs), (C, DD, NOWs, MMMF), where C is currency balances, DD is demand deposits, NOWs are negotiable orders of withdrawal[8] and MMMF are money market mutual funds. Hence, Belongia and Chalfant's results offered a choice among various levels of acceptable monetary aggregates for further analysis. This multiplicity of weakly separable groups is a common finding, so that other criteria, such as the Friedman one in Section 7.2 and others discussed later in this chapter, are further needed to select the most useful aggregate from among them.

7.4 Simple sum monetary aggregates

We are now going to switch from the theory of aggregation specified in the preceding section to the practical construction of the usual monetary aggregates, which are aggregations of the nominal values (X_i)[9] rather than the real values (x_i) of the assets. However, note that the assets included in any such aggregates must meet the separability criterion for aggregation.

8 NOWs are essentially a type of checkable savings deposits that can be transferred by negotiable orders of withdrawal corresponding to checks, so that NOWs are really a form of checkable deposits.

9 We use here the symbol X_i for the nominal value of the asset i even though it was used as the name of the asset in the preceding section.

In defining money, the most common *functional* form for the monetary aggregate is the *simple-sum aggregate* given by:

$$M = X_1 + \Sigma_i a_i X_i \quad i = 2, 3, \ldots m, a_i = (1, 0) \tag{7}$$

where:

M = nominal value of the monetary aggregate
X_1 = M1 (currency in the hands of the public plus demand deposits in commercial banks)
X_i = nominal value of the ith liquid asset

In Friedman's analysis (Friedman and Schwartz, 1963 a,b), the term a_i took the value 1 if inclusion of the ith asset gave a better result in explaining the level of national income than if it were excluded. However, the general functional form (7) was not exclusive to Friedman but was the most common form used in the 1950s and 1960s, and is still the most common monetary aggregation function in monetary economics.

A monetary aggregate given by (7) a priori assumes that:

(i) The coefficients a_i can take only the value zero or one so that all other values are excluded. Further, there is one-to-one substitution between the included assets.
(ii) An *infinite elasticity of substitution* exists among the assets with a non-zero coefficient, so that the included assets are perfect substitutes.

A generalization of (7) is a weighted sum aggregate that allows the coefficients a_i to take on any positive weights between zero and one. In this case, the monetary aggregate, designated as M', is given by:

$$M' = X_1 + \Sigma_i \, b_i \, X_i \quad i = 2, 3, \ldots, m \tag{8}$$

where, now, $0 \leq b_i \leq 1$. The weight a_i is sometimes called the *degree of moneyness* of asset i and can be specified a priori or on the basis of a statistical procedure. To illustrate the latter in line with Friedman's procedure for defining money, the values of the coefficients can be determined empirically by estimating the equation:

$$Y_t = a_0 + \Sigma_{i \, bi} \, X_{it} + \mu_t \quad i = 1, 2, \ldots, m \tag{9}$$

where Y is nominal national expenditure, μ is a stochastic term and $0 \leq b_i \leq 1$. (9) still defines money as that variable that "best" explains national expenditures, and the weights a_i of the assets are derived by multiple regression. The coefficient of each additional asset should decrease as the definition of money is broadened to include assets in a decreasing order of degree of moneyness. However, there are other criteria besides Friedman's for selecting among monetary aggregates, and extension of the definition to (9) is not common.

While (7) and (8) both specify simple sum aggregates, this term is usually associated with (7). It will henceforth be used in this sense. Many economists object to its underlying assumption that the elasticity of substitution must be either 0 (between an included and an excluded asset) or infinity (between any pair of the included assets). Two specifications of the monetary aggregate that use other elasticities are the variable elasticity of substitution function and the Divisia aggregates. These are presented in the next two sections.

7.5 The variable elasticity of substitution and near-monies

Chetty (1969) proposed using the following variable elasticity of substitution (VES) function for monetary aggregation. The VES function for the *nominal* value of the monetary aggregate M has the form:

$$M(X_1, \ldots, X_m) = \left(\Sigma_{i=1}^{m} a_i X_i^{1+v_i} \right)^{\frac{1}{(1+v_1)}} \tag{10}$$[10]

where X_i is the nominal quantity of the ith asset. In the special case where $v_i = v$ for all i, the VES function becomes the constant elasticity of substitution (CES) function, which is often used in the empirical estimation of the production function in microeconomics.[11] In this case, there would be an identical elasticity of substitution equal to $(-1/v)$ between each pair of assets. However, in our context, the assumption of $v_i = v$ would be an unjustified prior constraint on the data since, in the general case, the elasticity of substitution is likely to differ between different pairs of assets and is likely to vary over time for any given pair, so that v_i is unlikely to be identical for all i.

The partial elasticity of substitution[12] $\sigma_{i,j}$ is the elasticity of X_i/X_j with respect to the marginal rate of substitution (MRS) between them. It has the general definition:

$$\sigma_{i,j} = \frac{\mathrm{d}\ln(X_i/X_j)}{\mathrm{d}\ln(M_i/M_j)} \tag{11}$$

$$= \frac{(M_i/M_j)}{(X_i/X_j)} \frac{\mathrm{d}(X_i/X_j)}{\mathrm{d}(M_i/M_j)} \tag{12}$$

where $M_i = \partial M/\partial X_i$. The partial elasticity of substitution varies along any given indifference curve (or isoquant) – except for the case where $v_i = v$ for all i – and may differ between pairs of assets. It may be negative for any particular pair or pairs of assets, though (positive) substitution must dominate among all the assets taken together.

For (10), $\sigma_{i,1}$ between asset 1 and i varies with the quantities of assets and therefore with the period chosen. The average value of this elasticity over a sample period is calculated using the average values of the quantities of the assets, and is given by:

$$\bar{\sigma}_{i,1} = \frac{1}{-v_1 + (v_i - v_1)/\left[1 + \frac{a_i(1+v_i)}{a_1(1+v_1)} \frac{\overline{X_i}^{1+v_i}}{\overline{X_1}^{1+v_1}} \right]} \tag{13}$$

where the bar over a symbol indicates its average value.

Equation (10) has the nature of a utility (or production) function over the various monetary assets. The first asset is generally chosen to be M1,[13] with the other assets being near-money

10 The usual format of this equation can be obtained by setting $(1 + v_i)$ equal to $-\rho_i$.

11 Note that $v = -1$ turns the CES function into the Cobb–Douglas Function which is log-linear.

12 Note that the elasticity of substitution is different from the own-price elasticities of demand for the ith asset. The latter is defined as $\partial x_i/\partial p_i$. It is also different from the cross-price elasticity of demand for an asset i with respect to the price of another asset j, measured as $\partial x_i/\partial p_j$. The relationship between the elasticity of substitution and the own- and cross-elasticities of demand is discussed later in this section.

13 However, in some cases, the appropriate asset to start with can be currency in the hands of the public, with demand deposits considered to be a separate asset. This may well be preferable for some financially underdeveloped economies with few bank branches, often confined to the big cities and well-off individuals, so that

or other liquid assets. The monetary aggregate defined by (10) is sometimes referred to as *moneyness* or *liquidity*. To derive its estimating equations, assume that the individual maximizes (10) subject to the budget constraint:

$$\Sigma_i p_i X_i = A \tag{14}$$

where:

p_i = price of the ith asset

A = expenditures to be allocated over M1 and the other monetary assets.

Maximizing (10) subject to (14) gives the following first-order conditions, with λ as the Lagrangian multiplier:

$$M_i - \lambda p_i = 0 \quad \text{where } M_i = \partial M / \partial X_i, i = 1, 2, \ldots, m \tag{15}$$
$$\Sigma_i p_i X_i = A$$

which yield, for $i = 2, 3, \ldots$,

$$\frac{a_i(1+v_i)X_i^{v_i}}{a_1(1+v_1)X_1^{v_1}} = \frac{p_i}{p_1} \tag{16}$$

$$\ln X_i = \alpha_i + \frac{1}{v_i}\ln\frac{p_i}{p_1} + \frac{v_1}{v_i}\ln X_1 \tag{17}$$

$$\text{where } \alpha_i = \frac{1}{v_i}\ln\frac{a_1(1+v_1)}{a_i(1+v_i)}$$

Hence, for $i = 1, 2, \ldots, m$, the $(m-1)$ estimating equations for the i assets are of the form:

$$\ln X_i = \alpha_i + \frac{1}{v_i}\ln\frac{p_i}{p_1} + \frac{v_1}{v_i}\ln X_1 + \mu_i \tag{16'}$$

For $i > 2$, these equations need to be estimated by simultaneous regression techniques so as to meet the cross-equation restrictions on v_1. However, it is common to estimate (16′) by single equation regression techniques, as is done by Chetty. Such estimation provides the estimated values of v_i from the coefficient on $\ln(p_i/p_1)$. This value and the estimated value of (v_1/v_i) from the coefficient on $\ln X_1$ yield the estimate for v_1. From these values, the estimated values of the average partial elasticities of substitution can be calculated, as given by (13).[14]

An important theoretical and empirical issue in the above estimation is that of the measurement of the prices of financial assets. Viewed from the perspective of monetary assets as durable goods (which last beyond the current period), there are two

demand deposits do not constitute a medium of payments for much of the economy. Gebregiorgis and Handa (2005) report that for Nigeria, over the sample period 1970 to 2000, currency did better than other monetary aggregates such as M1 and M2.

14 One problem that often occurs in VES estimation is that the estimated coefficients (v_1/v_i) and $(1/v_i)$ respectively of $\ln X_i$ and $\ln(p_i/p_1)$ tend to be quite small, so that the calculation of v_i is very sensitive, as is that of the elasticity of substitution, to small differences in the estimated coefficients.

alternative ways of defining their prices. One of these is the purchase price, while its alternative is the user cost of the services rendered by the durable good during the period. Chetty (1969) used the former.[15] Defining a unit of the ith asset as having a value of unity at the end of the period, he specified its purchase price (i.e. its present discounted value) at the beginning of the period as $1/(1+R_i)$, where R_i is the nominal rate of return on the asset over the period. Hence, Chetty's study specified p_i as:

$$p_i = 1/(1 + R_i) \tag{18}$$

Designating the non-interest-paying asset 1 as M1 with $R_1 = 0$ implies that $p_1 = 1$, so that:

$$p_i/p_1 = 1/(1 + R_i) \quad i = 2, 3, \ldots \tag{19}$$

Empirical findings

We use Chetty's study for the United States for the period 1945–66 to provide one set of findings based on (17) and (19). This study considered four assets, with $i = 1, \ldots, 4$, and used their nominal values to estimate the nominal value M' of the monetary aggregate. The four assets in this study were:

X_1 = currency plus demand deposits (M1) in commercial banks (M1)
X_2 = savings/time deposits in commercial banks (TD)
X_3 = savings/time deposits in mutual savings banks (TDM)
X_4 = savings and loan associations shares (SLS).

Chetty's estimates of their elasticities of substitution, designated as $\sigma_{1,i}$, between money and the ith assets, were:

$$\sigma_{1,2} = \sigma_{\text{M1,TD}} = 30.864$$

$$\sigma_{1,3} = \sigma_{\text{M1,TDM}} = 35.461$$

$$\sigma_{1,4} = \sigma_{\text{M1,SLS}} = 23.310$$

Chetty interpreted these magnitudes as indicating that all three near-money assets were good substitutes for money. The estimated form of the monetary aggregate M was:

$$\hat{M} = \left[X_1^{0.954} + 1.020X_2^{0.975} + 0.880X_3^{0.959} + 0.615X_4^{0.981}\right]^{1.026} \tag{20}$$

Since the exponents of the variables on the right-hand side of (20) were close to one, as is the coefficient of X_2, M was approximated by Chetty as:

$$\hat{M} = X_1 + X_2 + 0.880X_3 + 0.615X_4 \tag{21}$$

15 This definition of the price of an asset is not based on its user or rental cost, discussed in Section 7.6 below. More recent contributions in the literature have not used Chetty's specification of prices but have substituted a user cost definition in its place.

The yield R_M on this monetary aggregate can be calculated as a correspondingly weighted index of interest rates on the near-money assets, so that it was specified as:

$$R_M = (R_2 + 0.880R_3 + 0.615R_4)/(1 + 0.88 + 0.615) \tag{22}$$

Chetty's estimation of the elasticities of substitution, by his single equation estimation and by his use of the cost rather than the user cost of assets, though limited in its usefulness by the form of the aggregation function specified a priori as (10), is valuable for providing an empirical procedure for estimating the degree of substitution between money and near-money assets. His results basically confirmed the general perception by economists in the 1960s of the close substitution between money and several other financial assets, and supported the evidence derived from estimates of the money demand function. Note that the financial deregulation and the considerable technological innovation in the financial sector in the USA since the 1960s have changed the elasticities of substitution from those reported by Chetty and have also created many new near-money assets, so that the monetary aggregate implied by estimating (17) would now be different from that derived by Chetty.

The VES (or CES) function estimates the elasticities of substitution which directly reflect the degree of substitution between money and near-money assets and are directly related to the debate on the appropriate definition of money, while the empirical studies of the *demand for money function* estimate the *own- and cross-interest elasticities of the demand for money*. The estimated values of the own- and cross-price elasticities are usually less than one, which are of a different order of magnitude from the elasticities of substitution reported by Chetty, so that we need a procedure for comparing the two concepts.

Comparison of elasticities of substitution with price elasticities

The elasticities of substitution are not directly comparable with price elasticities but are a component of the latter, and the two are difficult to compare for the general case. However, Feige and Pearce (1977, p. 461) reported the following relationship for the two-asset case, of which the first one is M1 with $R_1 = 0$.

$$\sigma_{1,2} = \left(\frac{1 + R_2}{R_2}\right)\left[E_{1,2} - E_{2,2}\right] \tag{23}$$

where:

$\sigma_{1,2}$ = elasticity of substitution between asset 1 and asset 2
$E_{2,2}$ = own-price elasticity of asset 2 (with respect to its own price)
$E_{1,2}$ = cross-price elasticity of asset 1 with respect to the return on asset 2
R_2 = return on asset 2.

To illustrate the reconciliation between the estimated large elasticities of substitution and the less-than-one price elasticities, let $i = 1$ for M1 and $j = 2$ for a near-money asset. Then, if $E_{1,2} = -0.4$, $E_{2,2} = 1.0$ and $R_2 = 0.04$, (23) implies that $\sigma_{1,2} = 36.4$, which is close to the elasticities of substitution reported by Chetty (Feige and Pearce, 1977, p. 460). Hence, small values of own- and cross-price elasticities can be consistent with very large elasticities of substitution, so that Chetty's estimates of the latter are not necessarily inconsistent with the own- and cross-elasticities usually occurring in the estimated demand functions.

7.6 User cost of assets

Chetty had specified a unit of the *i*th asset as having a *terminal* value of \$1 at the end of the period and its current price – equal to its present discounted value – as $[1/(1+R_i)]$, where R_i is the nominal rate of return on the *i*th asset itself. Similarly, the price per unit of the totally illiquid asset yielding R^* per period would be $[1/(1+R^*)]$, so that the relative price of the *i*th asset to that of the illiquid one would be $[(1+R_i)/(1+R^*)]$.

However, the proper concept for the *usage* of a durable good is its user cost for the services provided by it during the period (Barnett *et al.*, 1984). Comparing the *i*th somewhat liquid asset with the illiquid asset, the cost of using the liquidity of the *i*th asset during the period would be the return foregone by holding it rather than the illiquid asset. This foregone return per dollar invested in the *i*th asset is $(R^* - R_i)$ at the end of the current period. Discounting this return to the present gives the nominal and real (per dollar) user costs of the *i*th asset as:

$$\gamma_{it} = \frac{p_{it}(R^*_t - R_{it})}{1 + R^*_t} \tag{24}$$

$$\gamma^*_{it} = \frac{(R^*_t - R_{it})}{1 + R^*_t} \tag{25}$$

where:

γ_{it} = nominal user cost of *i*th asset in period *t*
γ^*_{it} = real user cost per dollar of the *i*th asset in period *t*
p_{it} = price of the *i*th asset in period *t*
R_{it} = nominal rate of return on the *i*th asset in period *t*
R^*_t = nominal rate of return on the totally illiquid asset

Further analysis in the body of this chapter assumes a zero tax rate, as in (24) and (25). The appendix to this chapter presents the corresponding equations if there is a tax on interest income.

The above measures of the user cost of the liquidity services provided by assets assume that the differential in their interest rates arises only from differences in their liquidity services. This will not be accurate unless the rates are determined by the market under perfectly competitive conditions and there are no other implicit or explicit charges on them – and the assets do not yield any services other than liquidity. Market rates usually do not satisfy these conditions. The market rates on assets may reflect differences in their associated services, such as investment advice, overdraft facilities, etc., other than their liquidity. Alternatively, some of the charges for these liquidity services may be through fixed charges and conditions – such as requirements for minimum balances, payment of interest only on minimum monthly balances, set-up charges and monthly service charges, – in addition to the differential in interest rates. Further, for investors, there may also be non-monetary personal "brokerage costs" as well as portfolio adjustment costs for reallocations among assets, or imperfect information about interest rates. These are not fully captured by the interest rate differentials.

Note that the user cost functions defined by (24) or (25) can be used with Chetty's variable elasticity of substitution function, as in Gebregiorgis and Handa (2005) and Lebi and Handa (2007), since Chetty's definition of the relative cost of assets was only a subsidiary hypothesis to his main hypothesis – that is, using a VES function for the aggregator function – and is not an integral part of it.

7.7 Index number theory and Divisia aggregates

Another approach to monetary aggregation is based on statistical index number theory and focuses on quantity and price data, rather than on utility or production functions, emphasizing the desirable properties of indices.[16] Among the statistically desired properties of an index number are that any changes in the prices of the components of the index change only the price index and any changes in the quantities of the components change only the quantity index, while the multiple of the price and the quantity indices thus computed equals the index of the expenditures on the services of the assets. The simple-sum aggregates do not meet several of these properties. One aggregate that does meet more of these properties is the Divisia aggregate, first proposed by François Divisia in 1925. The Divisia quantity and price indices possess the desired properties and make it tempting to select the Divisia quantity aggregate as the appropriate aggregator function for monetary assets.[17] The development and popularization of the (Törnqvist–Theil) Divisia monetary aggregate was initiated by Barnett (1980), who proposed the chain-weighted functional form, discussed later in this section, of this aggregate.

The Divisia quantity aggregate $x_t(x_{1t}, \ldots, x_{mt})$ for the m monetary assets for the period t is given by:

$$x_t(x_{1t}, \ldots, x_{mt}) = \prod_{i=1}^{m} x_{it}^{s_{it}} \tag{26}$$

where:

x_t = Divisia aggregate for period t
s_{it} = share of the ith asset in the expenditure on liquidity services in period t
$\prod_{i=1}^{m}$ = product, from 1 to m.

On the general nature of the Divisia index, (26) specifies the Divisia quantity aggregate as a CES (constant elasticity of substitution) function in which the elasticity of substitution is constant and identical at unity within each period.[18] Therefore, the functional form of the Divisia quantity aggregate x_t, as specified by (26), is the Cobb–Douglas one, so that it is essentially a weighted log-linear sum of the component assets, as in:

$$\ln x_t = \Sigma_i s_{it} \ln x_{it} \quad i = 1, \ldots, m \tag{27}$$

Further, (26) implies the computationally appealing growth equation:

$$\dot{x}_t = \sum_{i=1}^{m} s_{it} \dot{x}_{it} \tag{28}$$

16 Irving Fisher presented in the 1920s a detailed analysis of index numbers and their properties. Among the index numbers specified by him were the Laspeyres, the Paasche and Fisher's Ideal index, the last one being a geometric average of the former two.

17 Divisia aggregates are part of a class of statistical index numbers which are sometimes designated as "superlative."

18 Note in this connection that, for $v_i = v = -1$ for all i, the VES function in (10) becomes the Cobb–Douglas one.

where the dot on a variable indicates its growth rate. (28) shows the very attractive feature of Divisia aggregates that the growth rate of the aggregate is the sum of the weighted growth rates of the individual assets, with the weights being the shares of the total expenditures on liquidity services. In (26) to (28), the expenditure shares s_{it} could be held constant at s_i or allowed to change over time. The latter method allows the resulting Divisia index to incorporate new assets and to capture, to some extent at least, the impact of innovations in the financial sector.

To examine the ith asset's share of total expenditures, start with the nominal user cost of the ith asset as specified by (24). The nominal expenditure on the services of the ith asset is:

$$\gamma_{it} x_{it} = \frac{x_{it} p_{it} (R^*_t - R_{it})}{(1 + R^*_t)} \tag{29}$$

and the share of the expenditure on the ith asset out of the total expenditures on all assets is:

$$s_{it} = \frac{x_{it} p_{it} \dfrac{(R^*_t - R_{it})}{(1 + R^*_t)}}{\displaystyle\sum_{i=1}^{m} x_{it} p_{it} \frac{(R^*_t - R_{it})}{(1 + R^*_t)}} \tag{30}$$

$$= \frac{x_{it} p_{it} (R^*_t - R_{it})}{\displaystyle\sum_{i=1}^{m} x_{it} p_{it} (R^*_t - R_{it})} \tag{31}$$

As mentioned earlier, among the appealing features of the Divisia aggregates is that the weighting used for each asset is its share of the total expenditures on the flow of liquidity services provided by it.[19]

The chain-weighted Divisia index

Since the expenditure shares of liquid assets tend to differ across periods, the Divisia index is usually calculated in the following "chain-weighted" form:

$$x_t = x_{t-1} \prod_{i=1}^{m} \left[\frac{x_{it}}{x_{it-1}} \right]_t^{s^*_{it}} \tag{32}$$

where $s^*_{it} = 1/2(s_{it} + s_{it-1})$.

In these equations, the relevant weights in period t are s^*_{it}, which is the average share over periods t and $t-1$ of the expenditures on the ith asset. If expenditure shares shift over time, the time-linked weight s^* for period t will differ from those for other periods, since

19 Different assets provide different levels of liquidity services. Thus, currency and demand deposits are more liquid than long-term bonds. An aggregate measuring the total flow of these services should weight the holdings of the assets by a measure of the services provided by the respective asset, so that currency and demand deposits should have a higher weight than long-term bonds. The variable elasticity of substitution function and the Divisia aggregate do so, as do the weighted sum aggregates. However, the simple sum aggregates do not do so.

for instance, s^*_{it} will be the average of the expenditure shares s_{it} and s_{it-1}, while s^*_{it+1} will be the average of the expenditure shares s_{it+1} and s_{it}. The log-linear version of this Divisia quantity index provides the rate of change between periods t and $t-1$ as:

$$\ln x_t - \ln x_{t-1} = \Sigma_i s^*_{it}(\ln x_{it} - \ln x_{it-1}) \tag{33}$$

Since the resulting index takes account of changes in expenditures over time on the liquid assets, it will capture the impact of those innovations that alter the relative liquidity and demands of the component assets, which makes it preferable to an index that uses constant expenditure shares.

Economic theory and the appropriate form of aggregation

From an economic theoretic viewpoint, the appropriate form of the aggregation function is that which replicates actual economic behavior, whether or not it suits mathematical convenience or the desirable statistical properties of indices. Monetary theory does not prescribe this form, so it has to be determined by the data itself. There is no a priori reason why the data need necessarily behave according to the Divisia format, which requires a constant unit elasticity of substitution within a given period between each pair of its component assets. To illustrate this point for a financially developed economy, we expect, on the basis of our a priori knowledge, that the component assets (currency and demand deposits) of M1 possess high elasticities of substitution between them. This would make the Divisia aggregate of currency and demand deposits a poor approximation to the actual aggregate underlying the data. One way around this criticism is to rely on our a priori intuitive knowledge to combine simple sum aggregation and Divisia aggregation. In line with this, for financially developed economies, the Divisia aggregate is often *constructed by using M1 as its most liquid asset, where M1 is the simple sum aggregate* of currency in the hands of the public plus demand deposits in commercial banks. The unit elasticity assumption is then imposed between M1 and each of the other distinct assets, such as savings deposits, to construct a Divisia aggregate. But if savings deposits are almost perfect substitutes for M1, then M2 (= M1 + S) will be the primary asset in the construction of the Divisia index. This mode of construction is appealing because it uses common sense to blend the various convenient forms of aggregation. However, we still cannot be sure that this construct rather than another alternative, possibly with a different elasticity of substitution than unity, is the most appropriate one between M2 and other monetary assets for a given data set.[20] It is therefore necessary to construct and use appropriate statistical tests to judge the relative usefulness of the various aggregates. Such tests are presented later in this chapter.

7.8 The certainty equivalence monetary aggregate

Rotemberg (1991) and Rotemberg *et al.* (1995) proposed what they called the "currency equivalence" monetary aggregate (CEM) over the monetary assets. This too is a time-varying weighted average of the component assets but differs from the Divisia aggregates in that it

20 For instance, if the data is generated by a scenario where the elasticities of substitution are considerably in excess of unity for some pairs of assets and zero among others.

uses somewhat different weights from the latter. The functional form of the CEM aggregate for the weakly separable set of monetary assets is:

$$\ln \text{CEM}_t = \sum_{i=1}^{m} \theta_{it} \ln x_{it} \quad \theta_{it} = (R^*_t - R_{it})/R^*_t. \tag{34}$$

where x_i is the quantity of the ith asset, R^* is the nominal yield on a totally illiquid asset and R_{it} is the yield on asset i. The CEM index is log-linear, as is the Divisia one, so that it incorporates the assumption of identical elasticity of substitution equal to unity between each pair of assets. However, the CE index specifies θ_{it} as $(R^*_t - R_{it})/R^*_t$, while the Divisia aggregate, quite appropriately, defines user cost by $(R^*_t - R_{it})/(1 + R^*_t)$. The latter is preferable since its denominator is the appropriate mode of discounting from the end of the period to its beginning. However, if θ_{it} $(= (R^*_t - R_{it})/R^*_t)$ is viewed as an approximation of user cost, it becomes similar to the Divisia index, but with the property that the asset (presumably, currency) with $R_{it} = 0$ will have a weight of unity.[21]

The Divisia index assigns expenditure shares s_i (or the time-variant s_{it}), which must sum to unity, while the CE index assigns weights[22] which do not have to sum to unity.[23] The CE aggregate has the property that as long as currency (and zero-interest demand deposits) does not pay interest, its amount has a weight of one. Therefore, by normalizing the liquidity weight of currency (and demand deposits) at unity,[24] the interpretation of the CEM aggregate is that it specifies the amount of currency that would yield the same liquidity services as the assets in the monetary aggregate. With time-varying weights, just as for the chain-weighted Divisia aggregates, the CEM aggregate also adapts easily over time to reflect changes in the payments environment. While the performance of the CEM aggregate relative to its alternatives remains an open empirical question, the simple-sum and Divisia aggregates remain the more common modes of monetary aggregation.

7.9 Judging among the monetary aggregates

While the economist can decide on the appropriate level of aggregation on an a priori basis, this is often not a satisfactory procedure for empirical applications since the empirical characteristics of assets vary over periods and economies. Monetary economics therefore has a variety of empirical criteria for selecting among various aggregation procedures. One of these, the Friedman test, has already been discussed earlier in this chapter. In the following, we assume that weak separability tests were performed to determine the assets that have

21 Serletis and Molik (1999, p. 106) describe Rotemberg's CE index as a "(logarithmic) simple-sum index with a simple weighting mechanism added." However, it seems preferable to view the CE index as an approximation to the Divisia one, since, within a period, it shares with the latter the property of unit elasticity of substitution between pairs of assets and allows time-varying weights.

22 At the practical level of computing user costs, Rotemberg recommended using the return on common stocks as the benchmark rate – as opposed to using a long-term bond, which Barnett does – since all bonds offer some degree of liquidity services. This may bias the calculations on the user cost and expenditure shares.

23 Given the uncertainty of the yields on assets, both CE and Divisia indices should properly use the expected (rather than actual) returns in deriving user costs and expenditure shares.

24 By comparison, their weights in a Divisia index would be very much smaller than one.

been included in the aggregate, and that the choice that has to be exercised is between the functional forms of the aggregate.

7.9.1 Stability of the money demand function

If the estimated demand function for a monetary aggregate is to be useful for prediction, it should have a high R^2 and be stable over different sample periods. If the function is not stable, its value for predicting money demand in future periods and for policy purposes is limited. While this requirement may not seem to be very stringent, it is not always satisfied and rejects many of the estimated demand functions. For instance, the money-demand functions estimated for 1980s and 1990s for most of the commonly used monetary aggregates show a high degree of instability for many countries.

7.9.2 Controllability of the monetary aggregate and policy instruments and targets

If the central bank is to consider the monetary aggregate useful for policy purposes, the bank must be able to control it through the policy instruments at its disposal. Assuming that the central bank uses the nominal monetary base M0 as its control instrument, it will be concerned with the relationship between M0 and the monetary aggregate M. A simple linear relationship between M and M0 is given by:

$$M = a_0 + a_1 \mathrm{M0} \tag{35}$$

where:

M = nominal value of the monetary aggregate
M0 = nominal value of the monetary base
a_1 = monetary base-money multiplier ($\partial M / \partial \mathrm{M0}$)

The central bank can control M through M0 only if a_0 and a_1 are stable. Different monetary aggregates possess different values of these coefficients with different degrees of stability. The preferred aggregate would be one with stable values of the coefficients and a high R^2 in estimations.

Note that this criterion becomes irrelevant if the central bank does not believe that its manipulation of monetary aggregates confers a predictable benefit in terms of aggregate demand and its final goal variables, such as output and unemployment. In fact, many central banks now hold the view that manipulation of the interest rate provides better control over these variables, so that there is a general tendency to downplay the relevance of monetary aggregates (see Chapter 13).

7.9.3 Causality from the monetary aggregate to income

Take a policy instrument X, which could be a monetary aggregate or an interest rate. To be useful for controlling nominal national income Y, changes in X should cause changes in Y. The statistical procedure for determining causation between variables is the Granger–Sims test for causality. This procedure is a statistical determination of causality and judges (statistical) causality to go from a variable X to Y – that is, a change in X causes a change in Y – if the data on them shows a lag in the impact of X on Y. That is, if the *lagged* values of X are significant in a regression of Y on X, X is said to Granger-cause Y. If the expected *future*

values of X are significant, Y is said to Granger-cause X.[25] If both the lagged and the future values of X are significant, Granger causality runs two-way between X and Y. Hence, lags and/or leads are essential for this procedure to give any results on causality; this procedure does not detect causality if there are no leads or lags but only a *contemporaneous* impact from X to Y, or from Y to X.

Sims (1972), in an early application of the Granger procedure for statistical causality to the relationship between money and income, defined *one-way statistical causality thus*:

> If and only if causality runs one way from current and past values of some list of exogenous variables to a given endogenous variable, then in a regression of the endogenous variable on past, current, and future values of the exogenous variables, the future values of the exogenous variables should have zero coefficients.
>
> (Sims, 1972, p. 541).

This test relies on regression analysis to determine the pattern of lags among the variables. If nominal national income Y is the endogenous variable and the exogenous variables are the policy instrument X and other variables Z, the relevant regression equation is of the form:

$$Y_t = \Sigma_i a_i X_{t-i} + \Sigma_i b_i Z_{t-i} + \Sigma_i \gamma_i Y_{t-i} + \Sigma_j c_j X_{t+j} + \mu_t \tag{36}$$

where Z is the vector of variables other than X and Y and μ is the error term. Note that (36) includes not only the current and lagged values of X, Z and Y but also some future values of X through the term $\Sigma_j c_j X_{t+j}$. If the estimated coefficients of X are statistically different from zero but those of its future terms are zero, then one-way Granger causality is said to exist from X to Y.

Usually a similar regression is also run with X as the dependent variable and Y among the independent ones. Such a regression would be of the form:

$$X_t = \Sigma_i \alpha_i Y_{t-i} + \Sigma_i \beta_i Z_{t-i} + \Sigma_i \lambda_i X_{t-i} + \Sigma_j k_j Y_{t+j} + \mu_t \tag{37}$$

For this regression, with X on the left side, one-way causality from Y to X requires that the coefficients of the lagged values of Y be significantly different from zero, while those of its future values be zero.

Two-way causality between X and Y exists in (36) if *both* the lagged and the future terms in X have non-zero coefficients. This can be verified from (37) if both the lagged and future terms in Y have non-zero coefficients. We illustrate these remarks and the nature of the Granger causality test by a simple example where:

$$Y_t = a_1 X_t + a_2 X_{t-1} + a_3 X_{t+1} + \mu_t \tag{38}$$

25 If economic theory implies such an impact, the following specification of this procedure will lead to erroneous conclusions on the direction of causality.

Let a_2 be finite with a value other than zero. Under the above procedure, X is taken to cause Y if $\hat{a}_2$ (the estimated value of a_2) $\neq 0$. Suppose we rewrite (38) as:

$$X_{t-1} = (1/a_2)Y_t - (a_1/a_2)X_t - (a_3/a_2)X_{t+1} - (1/a_2)\mu_t \tag{39}$$

which can be rewritten as:

$$X_t = (1/a_2)Y_{t+1} - (a_1/a_2)X_{t+1} - (a_3/a_2)X_{t+2} - (1/a_2)\mu_{t+1} \tag{40}$$

Now, treating X as the dependent variable and Y as the independent one of the regression, since a_2 is finite, $(1/a_2) \neq 0$. Hence the regression of X_t – as the dependent variable – on the future Y_{t+1} would yield a non-zero value of the coefficient of Y_{t+1}. Extending this argument to the general case shows that the non-zero coefficients of any of the future values of a variable indicate reverse causation from the dependent variable to that variable.[26] Note also that the coefficient of a contemporaneous term X_t does not provide any information on the direction of causality.

Applying these arguments back to the estimation of (36), which is a regression of Y_t on the X and Z variables, if the lagged values of X had some non-zero coefficients a_i, the implication would be that causation runs from X to Y. But if the estimated values of any of the coefficients c_j (of the future values of the money supply) were also non-zero, the implication would be that causation runs from Y to X. For one-way causation from X to Y, some a_i should be non-zero while all c_j should be zero, but for one-way causation from Y to X, some c_j should be non-zero while all a_i should be zero. For two-way causation between Y and X, some of each set of coefficients should be non-zero.

Using causality tests to judge among monetary aggregates and interest rates

Since it is not a priori always clear for a particular economy and data set whether the interest rate or the money supply is the truly exogenous policy instrument, the other being the endogenous one, causality tests should be run to determine the direction of causality between them.

The different monetary aggregates do not perform equally well in terms of one-way causality from money to nominal national income, so the Granger–Sims test can be used as a selection device among them. Note that this test cannot be used if there are no leads or lags in the relationships among the variables, as would be the case where there is only a contemporaneous relationship. In such a context, the following test provides some information on the direction of causality.

7.9.4 Information content of economic indicators

In the rare case where there are no leads or lags between monetary variables and income, so that Granger causality cannot be used to discriminate among the former, one test that can be

26 Note that the serial correlation of the error term in a regression can lead to spurious results: such serial correlation can cause the estimated coefficients of the future terms to erroneously differ from zero.

used is one that calculates the "information content" of the stochastic variable X for Y. For this test, the stochastic linear model used between Y and X is assumed to be:

$$Y_t = a_0 + a_1 X_t + \mu_t \tag{41}$$

where μ_t is white noise. The expected information content of X with respect to Y is measured by:

$$I(Y|X) = \frac{1}{2} \ln \left[\frac{1}{(1 - R^2)} \right] \tag{42}$$

where R^2 is the coefficient of determination. Note that this test uses only the contemporaneous values of the variables. A higher value of $I(Y|X)$ indicates a higher information content of X for Y, so that the form of X that has a higher value for this statistic would be preferable for explaining Y.

7.9.5 The St Louis monetarist equation

The reduced-form equation from the short-run macroeconomic IS–LM models with an exogenous money supply for the relationship between nominal national income Y and the money stock M is:

$$Y = f(\boldsymbol{M}, \boldsymbol{G}, \boldsymbol{Z}) \tag{43}$$

where:

Y = nominal national income
$\boldsymbol{M}$ = vector of past and present nominal values of the appropriate monetary aggregate
$\boldsymbol{G}$ = vector of past and present values of the appropriate fiscal variables
$\boldsymbol{Z}$ = vector of the other independent variables.

The linear or log-linear stochastic form of the above equation, popularized by the St Louis monetarist school (Anderson and Jordan, 1968), was:

$$Y_t = \Sigma_i a_i M_{t-i} + \Sigma_j b_j G_{t-j} + \Sigma_i c_i Z_{t-i} + \Sigma_i \gamma_i Y_{t-i} + \mu_t \tag{44}$$

Equation (44) is often called the St Louis equation and represents a popular method for determining the impact of monetary and fiscal variables on nominal income. It can be estimated in levels or first differences of the variables. There are many different ways of specifying the fiscal variable G. Among these is government expenditures and the fiscal deficit, or their values at full employment. The set of variables in Z also tends to vary between studies.

The monetary aggregate in (44) can be represented by one of its different forms. To make the choice among these forms for given versions of the Y, G and Z variables, (44) is estimated using different monetary aggregates. These estimates are examined for the signs, significance and stability of the coefficients. The monetary aggregate that performs "better" – in terms of the coefficient of determination, the plausibility of the estimated values of the coefficients and their stability – is considered preferable to the other aggregates.

Since many central banks now use an interest rate as their main policy instrument, the appropriate form of the St Louis equation in this context becomes:

$$Y_t = \Sigma_i a_i R_{t-i} + \Sigma_j b_j G_{t-j} + \Sigma_i c_i Z_{t-i} + \Sigma_i \gamma_i Y_{t-i} + \mu_t \tag{45}$$

where R is the relevant nominal interest rate. While this equation can be used to judge the interest rate most relevant to explaining Y, its estimation will not provide any information on the appropriate form of the monetary aggregate.

7.9.6 Comparing the evidence on Divisia versus simple-sum aggregation

Theoretical considerations

As explained earlier in this chapter, in a fairly static context, simple-sum aggregation between two monetary variables is relatively more appropriate – and therefore likely to perform "better" empirically – than their Divisia aggregate if the two are close substitutes, so that they possess a high elasticity of substitution. But their Divisia aggregate is relatively more appropriate, and likely to do better empirically, for any two assets that have a relatively low degree of substitution, especially if their elasticity of substitution is close to unity. The wider the existing degree of aggregation, the more likely is an additional asset to fit into the latter case, so that the broader the definition of the monetary aggregate, the more likely is the Divisia aggregate to perform better than its simple-sum counterpart.[27] Since the empirical findings depend upon the actual degree of elasticity of substitution among assets, a financial innovation that changes this elasticity could alter the results on the appropriate monetary aggregate. In fact, the past several decades have seen a considerable amount of innovation in financial intermediation and in payments technology, causing shifts in the estimated demand functions for the monetary aggregates and therefore also in the relationships between money and national income. This process has been accompanied by increases in the liquidity of many assets that are not in M1 and even those not in M2. Divisia aggregates with time-varying weights can capture such shifts through changes in these weights over time. Simple-sum aggregates in their standard form have a fixed weight of unity for each asset in the aggregate and zero for excluded assets, so that their rigidity in weights makes them inappropriate for a period of changing liquidity patterns. Therefore, for broad aggregates and for recent periods, the expectation is that simple-sum aggregation is likely to do worse than Divisia aggregation with time-variant weights.

However, while Divisia aggregates are expected to perform better – especially for very broad aggregates – from a theoretical and statistical perspective under a changing payments technology, the simple-sum aggregates retain their popularity for the public and the policy makers. The latter are easier to grasp and compute, whereas the former require intricate calculations. Further, if the Divisia aggregates are to be responsive to the changing

27 Note that if the rates of return on both currency and demand deposits are zero and their non-monetary costs are ignored, the user cost of the two would be identical. This would impart a degree of similarity in their weighting in Divisia aggregation. Barnett *et al.* (1984) argue that since the return on currency and demand deposits move similarly, their simple-sum aggregate M1 and its Divisia aggregate DM1 tend to perform similarly under the various criteria.

liquidity patterns, they need recalculation of the user costs and expenditure shares for each period.

Empirical findings on monetary aggregation

There are several excellent surveys of the empirical literature on monetary aggregates. Among these are Goldfield (1973), Judd and Scadding (1982), Rotemberg (1993), Belongia and Chrystal (1991), Chrystal and MacDonald (1994), Sriram (1999). We present below the results of just a few of these studies comparing simple-sum aggregates (SM) with Divisia ones (DM). Among these was that of Barnett *et al.* (1984), who used quarterly US data from 1959 to 1982 and conducted various tests related to money demand, velocity, and causality between money and income, using reduced-form (St Louis) equations for income. The authors reported that neither the simple-sum nor the Divisia aggregates uniformly performed better than the other for all the criteria considered and no one aggregate was uniformly the best one. The Divisia aggregates generally performed better than the simple-sum aggregates, except for M2, in causality tests. In the money-demand functions and in respect of stability, the Divisia aggregates fared better, but in the reduced-form equations for income, SM1 did better than DM1. Barnett and colleagues concluded that neither SM1 nor DM1 dominated the other under all criteria. However, the Divisia measures gave increasingly better results at higher levels of aggregation than the corresponding simple-sum measures.

We have already mentioned Belongia and Chalfant (1989) in connection with weak separability tests. They tested the Divisia and simple-sum versions of their weakly separable groups in St Louis equations and in equations relating the monetary aggregate to the monetary base, to capture the degree of controllability, for US quarterly data over 1976 to 1987. In the St Louis equation test there was a clear preference for M1 over broader measures. In the controllability tests, none of the measures did very well. Re-estimation for the period 1980 to 1987 led the authors to favor the Divisia over the simple-sum version of M1A. Belongia (1996) tested the relationship between monetary aggregates and nominal income for the USA for the period 1980:1 to 1992:4 for the simple-sum and Divisia aggregates. His finding for this relationship was that the Divisia aggregates performed better than the simple-sum ones and the instability of the money–income relationship was eliminated if the Divisia aggregates were used.

The above studies used standard estimation methods but not cointegration techniques. Although these are explained in the next chapter, we do want to cite in this chapter some results based on these techniques. Chrystal and MacDonald (1994) compared the simple-sum and Divisia aggregates for a number of countries, including the USA, UK and Canada, for various periods over the 1970s and 1980s. Their tests encompassed the St Louis equation and causality tests, and used cointegration. Note that the data for all the countries is contaminated by financial innovation over the sample period. We first examine their findings using the St Louis equation. For the USA, while M1 and M1A tended to perform better than their Divisia versions, the latter did better for broader definitions of money. But DM2 did not sufficiently dominate SM2 for the former to be clearly preferable. For the UK, the authors considered financial innovation to have sufficiently distorted the data to include only M0 – that is, the monetary base – and M4 in their estimations. While DM4 was clearly preferable to SM4, there was an inadequate basis for choosing DM4 over M0. For Canada, while SM1 was slightly preferable to DM1, the broader Divisia aggregates were preferred to their simple-sum versions.

For the causality tests, Chrystal and MacDonald considered it important to include an interest rate in their cointegration and error correction models. For the UK and Canada, both the Divisia and the simple-sum aggregates showed little causal impact on real output. For the USA, the Divisia measures were significant but not the simple-sum measures. The overall comparison, based on both the St Louis equation and the causality tests, did not, therefore, strongly favor one of the aggregates, though there was somewhat stronger support for the broad Divisia aggregates over the corresponding broad simple-sum aggregates. For the USA, Chrystal and MacDonald found that this was especially so after 1980, when financial innovation occurred at a faster rate. By comparison, pre-1980 US data did not favor the Divisia aggregates.[28]

Gebregiorgis and Handa (2005) estimate simple-sum, VES and Divisia aggregates, employing the user cost concept, for Nigeria for the period 1970:1 to 2000:4 and examine their relative performance, using cointegration analysis, for the determination of industrial production. Contrary to the usual findings for developed economies, this study reports that for Nigeria currency did as well as or better than any narrow- or broad-money measure. However, this need not be surprising since most of the population of Nigeria does not have access to or utilize banking facilities. Further, they report that simple-sum aggregates of M1 and M2 do better than their VES and Divisia counterparts. In fact, Divisia M2 does worse than currency and the simple-sum and VES aggregates for M1 and M2.[29]

Lebi and Handa (2007) study monetary aggregation at the M2 level for Canada, using the various tests outlined in this chapter to judge among them. Statistical problems that arose in its construction made it impossible to construct the VES aggregate for M2. Their finding is that the Divisia and currency equivalence M2 indices perform comparably well in most tests, with CEM2 marginally outperforming DVM2 in the information criterion and the St Louis equation tests. The simple-sum measure SSM2 dominates in the Granger causality tests between money and income, and also performs moderately well in the St Louis equation tests.[30] In general, the econometric estimates from the cointegration analysis for the money-demand functions do not indicate a clear winner as the monetary aggregate of choice. The authors conclude that no one monetary aggregate dominates all others in meeting each of the possible criteria.

To conclude, neither the simple-sum nor the Divisia aggregates do better than the other for all levels of aggregation, all relevant tests and all periods. There is usually no clear dominance of DM1 over SM1, and the two measures have very similar growth rates. But, in general, the broader the aggregate, the more likely is the Divisia measure to perform better. This is especially so for Divisia aggregates with time-varying weights. However, note

28 Serletis and Robb (1986) found a low degree of substitution among the monetary assets in Canada and little justification for M2 and broader aggregates. Their study favored the Divisia aggregates over the simple-sum aggregates.

29 They reasoned that among the segment of Nigeria's population, mostly in the major urban centers, that uses checking accounts on a regular basis, the elasticity of substitution between currency and checking deposits is likely to be very high, whereas it is likely to be closer to zero in the rural areas with limited banking facilities. While one cannot a priori predict the elasticity of substitution between currency and checking deposits in the aggregate data, a variable elasticity of substitution (VES) function should have a better likelihood of doing as well as or better than a Divisia one.

30 It should be noted that in cases of divergence between the results from the tests in the levels of the variables and their first differences, the latter are preferred.

that, as aggregation is broadened to include assets with very low substitution elasticities, further broadening of the Divisia aggregate makes little difference to the performance of the aggregate.

7.10 Current research and policy perspectives on monetary aggregation

As already mentioned at various points in this chapter, the demand for monetary aggregates has proved to be unstable in recent periods, largely because of innovations in the variety of monetary assets and modes of payment. This led a Governor of the Bank of Canada to remark at one point that the Bank did not abandon monetary aggregates, they abandoned central banks as an effective policy instrument. The result has been that many central banks in the developed economies have given up control over monetary aggregates in favor of interest rates as the main policy instrument for controlling aggregate demand.

In line with this development, monetary economists seem to have lost interest in the estimation of monetary aggregates, so that few significantly new studies have appeared on this topic since about 1990. Further, whatever limited interest there remains in monetary aggregates, the focus seems to have shifted from Divisia and other complex forms of aggregation back to simple forms, which are more readily understood by policy makers and the public.

Conclusions

The starting point for any aggregation is to establish weak separability among the assets which are to be included in the aggregate. Once this has been established, the next choice is to construct a simple-sum aggregate, a variable elasticity of substitution aggregate, and a Divisia or some other aggregate. However, the usual comparison in the literature is between the simple-sum aggregates and the Divisia aggregates, largely because of problems in reliable estimates of the variable elasticity of substitution.

Simple-sum aggregation assumes a priori infinite elasticities of substitution between each pair of the component assets of the aggregate, and Divisia aggregation assumes a priori unit elasticity between each pair of the component assets of its aggregate. Intuitively, the former will tend to be the relevant one if the empirical elasticities of substitution are high, and the latter if they are low. Thus the public's currency balances and demand deposits in developed financial sectors tend to be almost perfect substitutes and their simple-sum aggregate would be the more appropriate one. In fact, even the common constructs of the Divisia aggregates start with the simple-sum aggregate M1 for currency and demand deposits as the base aggregate. However, in less developed economies with poor banking facilities in the rural areas and with high brokerage costs of banking, the rural public's currency balances and demand deposits are likely to have limited elasticity of substitution, so that their Divisia aggregation may be more appropriate in such a context.

For the financially developed economies, the simple-sum M1, composed of currency in the hands of the public and demand deposits in commercial banks, is usually taken as one unified asset in constructing the broader Divisia aggregates. Further, since the return on M1 is usually taken to be zero, its rental cost would be higher than of any other asset

with a positive rate of return, giving it a correspondingly higher weight in the Divisia aggregate.

While the Divisia aggregates are superior in terms of desirable index number properties, their a priori assumption of the unit elasticity of substitution limits their usefulness for aggregation over assets that are very close substitutes – and therefore have elasticities of substitution that are substantially higher than unity. Intuitive knowledge of the financial markets strongly indicates that there exist several near-monies[31] that are close substitutes for money and would have numerically very high elasticities of substitution. In fact, both intuition and Chetty's procedure, though it poses problems in estimation, do indicate that many of the financial assets can have very high elasticities of substitution.

As the aggregation is broadened to bring in more and more assets into the aggregate, the additional assets are likely to have successively lower elasticities of substitution with currency balances and demand deposits, so that Divisia – as against simple-sum – aggregation becomes correspondingly more relevant to the further broadening of the aggregate. In general, there is no a priori procedure for deciding which aggregation procedure is better over various groups of assets.

In a dynamic context with innovation changing the liquidity weights, user costs and expenditure shares over time, the Divisia aggregates have the advantage that the relative shares or weights of the assets are allowed to vary over time, so that the computed index has a flexibility not available to simple-sum aggregates in capturing the impact of innovations. The former are therefore likely to perform better than the latter in a dynamic period of financial and technical innovation. However, they have not always done better than simple-sum aggregates in different types of empirical tests; so that there is no convincing evidence that the former unambiguously dominate over the latter. Given this lack of convincing evidence and the familiarity and ease of computation of simple-sum aggregates, these aggregates continue to maintain their popularity for the policy makers, the public and much of the economics profession. Sriram (1999) presents the usage of the three types of monetary aggregates in 38 studies that had employed cointegration techniques, separating them between those based on developed and on developing economies. Of these 38 studies surveyed, only one out of twelve for the developed economies had investigated Divisia aggregation, and only one out of twenty-four for the developing economies had done so.

Given that different aggregates tend to perform differently under various tests, we conclude that monetary aggregates need to be tailored to different criteria and purposes, so that the first step should be to choose the intended use of the monetary aggregate or, alternatively, to first establish a hierarchy among the criteria on some basis of desirability. Further, the profession might have to settle for the simultaneous use and acceptance of several different monetary aggregates, though for different purposes. This, in turn, suggests that the quest for a single all-purpose monetary aggregate should be abandoned. However, for central bank policy purposes and announcements, the most appropriate aggregates remain simple-sum ones, though at different levels of aggregation.[32]

31 For example, checkable savings deposits are now very close substitutes for demand deposits for most individuals.

32 Few central banks in the major developed economies now claim to use monetary aggregates as their main, or one of the main, policy instruments. One of the few that does so is the European Central Bank, which uses the simple-sum M3.

Appendix: Divisia aggregation

The symbols in this appendix have the following meanings:

- x Divisia index for the real value of the composite asset "money"
- x_i real value of asset X_i
- p Divisia index of the price level of the composite real asset "money"
- p_i price of asset X_i
- s_i share of the expenditures on X_i
- γ_i nominal user cost of X_i
- γ^*_i user cost per dollar of X_i
- R^* maximum available yield – called the "benchmark rate" – over the holding period
- R_i pre-tax nominal yield on the ith asset
- τ marginal tax rate
- Π product symbol

For Divisia price and quantity aggregates, we want to construct an index of expenditures, px, where:

$$px = \Sigma_i p_i x_i \tag{46}$$

The rate of change of px is given by:

$$\frac{1}{px}\frac{\mathrm{d}(px)}{\mathrm{d}t} = \frac{\mathrm{d}x}{x} + \frac{\mathrm{d}p}{p} = \frac{\sum_{i=1}^{m}(x_i \mathrm{d}p_i + p_i \mathrm{d}x_i)}{\sum_{i=1}^{m} p_i x_i}$$

$$= \sum_{i=1}^{m} s_i \left[\frac{\mathrm{d}p_i}{p_i} + \frac{\mathrm{d}x_i}{x_i}\right] \tag{47}$$

$$\text{where } s_i = \frac{p_i x_i}{\sum_{i=1}^{m} p_i x_i} \tag{48}$$

For the Divisia aggregates, we want to impose the conditions that the quantity index x in the overall index px changes if and only if the quantities (but not the prices) of the component assets change, and the price index p in the overall index changes if and only if there is a change in the prices (but not in the quantities) of the component assets. Formally, the required conditions are that:

$$\frac{\mathrm{d}x}{x} = \sum_{i=1}^{m} s_i \frac{\mathrm{d}x_i}{x_i} \tag{49}$$

$$\frac{\mathrm{d}p}{p} = \sum_{i=1}^{m} s_i \frac{\mathrm{d}p_i}{p_i} \tag{50}$$

Integrating (49) and (50) and using the base period prices and quantities respectively designated as p_0 and x_0 as the constants of integration gives:

$$x_t = x_0 \exp\left[\int_0^t \sum_{i=1}^{m} s_i \frac{\mathrm{d}x_i}{x_i}\right] \tag{51}$$

$$p_t = p_0 \exp\left[\int_0^t \sum_{i=1}^{m} s_i \frac{\mathrm{d}p_i}{p}\right] \tag{52}$$

Since economic variables are observed and measured in discrete rather than continuous time, the preceding continuous indices are approximated by the following discrete versions:

$$x_t = \prod_{i=1}^{m} x_{it}^{s_{it}} \tag{53}$$

$$\ln x_t = \sum_{i=1}^{m} s_{it} \ln x_{it} \tag{53'}$$

and the Divisia price index as:

$$p_t = \prod_{i=1}^{m} p_{it}^{s_{it}} \tag{54}$$

$$\ln p_t = \sum_{i=1}^{m} s_{it} \ln p_{it}$$

A comparison between two periods t and $t-1$, with $t-1$ treated as the base period for the index, changes (51) and (52) to:

$$x_t = x_{t-1} \exp\left[\int_{t-1}^{t} \sum_{i=1}^{m} s_i \frac{\mathrm{d}x_i}{x}\right] \tag{51'}$$

$$p_t = p_{t-1} \exp\left[\int_{t-1}^{t} \sum_{i=1}^{m} s_i \frac{\mathrm{d}p_i}{p}\right] \tag{52'}$$

The preceding indices are based on time-invariant expenditure shares. However, these change over time for a variety of reasons, including the emergence of new assets and changes in the liquidity of existing ones. Therefore, given the likelihood that the expenditure on each asset is not constant over time so that s_{it} and s_{it-1} may differ, equations (51′) and (52′) can be approximated by:

$$x_t = x_{t-1} \prod_{i=1}^{m} \left[\frac{x_{it}}{x_{it-1}}\right]_t^{s^*_{it}} \tag{55}$$

$$p_t = p_{t-1} \prod_{i=1}^{m} \left[\frac{p_{it}}{p_{it-1}} \right]^{s^*_{it}} \tag{56}$$

$$s^*_{it} = \frac{1}{2}(s_{it} + s_{it-1}) \tag{57}$$

Equations (55) and (56) yield the "chain-weighted" Divisia indices. In these equations, the relevant weights in period t are s^*_{it}, the average shares over periods t and $t-1$ of the expenditures on the ith asset. From (55) and (56), the rates of change in the Divisia quantity and price indices between periods t and $t-1$ are given by:

$$\ln x_t - \ln x_{t-1} = \Sigma_i s^*_{it}(\ln x_{it} - \ln x_{it-1}) \tag{58}$$

$$\ln p_t - \ln p_{t-1} = \Sigma_i s^*_{it}(\ln p_{it} - \ln p_{it-1}) \tag{59}$$

Measuring prices by the user costs of liquidity services

We now return to the general form of the Divisia index and introduce prices measured by user costs. The before-tax nominal user cost γ_{it} and the real user cost γ^*_{it} of the ith asset in period t are:

$$\gamma_{it} = \frac{p_{it}(R^*_t - R_{it})}{1 + R^*_t} \tag{60}$$

$$\gamma^*_{it} = \frac{(R^*_t - R_{it})}{1 + R^*_t} \tag{61}$$

Equation (60) specifies the present discounted value of the nominal user cost of holding the ith asset. As discussed in the body of the chapter, this user cost is defined for the flow of liquidity services provided by each asset, with these services assumed to be the only reason for the differences in their returns. The numerator $(R^*_t - R_{it})$ is the end-of-the-period *return foregone* from a dollar of the ith asset for its holding period;[33] R^*_t is the interest rate on the totally illiquid asset.[34] Therefore, the denominator $(1 + R^*_t)$ in (60) is the discount factor to convert the end-of-the-period value to its present (beginning-of-the-period) value using the return R^*_t on the most illiquid asset.[35] Equation (61) divides the nominal user cost by p_t to get the real user cost per dollar invested in the asset.

33 Both interest rates should be over the holding period of the ith asset. However, R_i is usually over a shorter holding period than R^*, which would be the interest rate on a long-term bond. To adjust for the difference in holding period, all rates have to be adjusted to the same maturity.

34 In practice, this asset is taken to be one with the highest interest rate, so that all other assets would have positive user costs.

35 Note that this designation (Barnett *et al*., 1984) of the price or cost of the ith asset differs from that in Chetty (1969). In the present notation and assuming a zero tax rate, Chetty's measure of the current price of the asset would have been $\{1/(1 + R_{it})\}$. It is the present discounted value of $1 obtained at the end of the period from the asset i, with the discount factor $(1 + R_{it})$.

As against these two measures, Donovan (1978) defined the price of the ith asset as $(1 + R_{it})/(1 + R^*_t)$, where $(1 + R_{it})$ is the amount that one dollar invested in asset X_i in period t would yield at the end of the period, and $(1 + R_{it})/(1 + R^*_t)$ is its present value.

The corresponding expenditure share on asset i is specified by:

$$s_{it} = \frac{x_{it}p_{it}\dfrac{(R^*_t - R_{it})}{(1+R^*_t)}}{\sum_{i=1}^{m} x_{it}p_{it}\dfrac{(R^*_t - R_{it})}{(1+R^*_t)}} \tag{62}$$

$$= \frac{x_{it}p_{it}(R^*_t - R_{it})}{\sum_{i=1}^{m} x_{it}p_{it}(R^*_t - R_{it})} \tag{63}$$

From (53′) and (63), the Divisia quantity index is given by:

$$\ln x_t = \sum_{i=1}^{m} \ln x_{it} \left[\frac{x_{it}p_{it}(R^*_t - R_{it})}{\sum_{i=1}^{n} x_{it}p_{it}(R^*_t - R_{it})} \right] \tag{64}$$

Adjustments for taxes on rates of return

If the tax authorities levy a constant tax rate of τ_t in period t on interest income but not on liquidity services, the corresponding equations for user cost, shares of expenditures on each asset and the Divisia quantity index adjusted for this constant tax rate would be as in the following equations:

$$\gamma_{it} = \frac{p_{it}(R^*_t - R_{it})(1-\tau_t)}{1+R^*_t(1-\tau_t)} \tag{65}$$

$$\gamma^*_{it} = \frac{(R^*_t - R_{it})(1-\tau_t)}{1+R^*_t(1-\tau_t)} \tag{66}$$

$$s_{it} = \frac{x_{it}p_{it}(R^*_t - R_{it})(1-\tau)}{\sum_{i=1}^{m} x_{it}p_{it}(R^*_t - R_{it})(1-\tau)} \tag{67}$$

$$\ln x_t = \sum_{i=1}^{m} \ln x_{it} \left[\frac{x_{it}p_{it}(R^*_t - R_{it})(1-\tau_t)}{\sum_{i=1}^{n} x_{it}p_{it}(R^*_t - R_{it})(1-\tau_t)} \right] \tag{68}$$

There are thus different ways of defining the price or user cost of an asset. (60) seems preferable to the other measures because of its focus in the numerator on the return foregone from holding a liquid asset as against an illiquid one. It also uses the market discount rate on illiquid assets that would be the relevant one for loans to the individual.

Summary of critical conclusions

- A simple-sum aggregate assumes perfect substitution – that is, infinite elasticity of substitution – among its components.
- The variable elasticity of substitution (VES) monetary aggregate allows the elasticity of substitution to differ between pairs of its components and does not a priori impose on each one a given elasticity of substitution. It allows the elasticity of substitution to lie in the range from zero to infinity.
- A Divisia aggregate assumes unit elasticity of substitution among its components. This aggregate weights each component asset by its share of expenditure out of the total expenditure on the aggregate. Its chain-weighted version allows these shares to change over time to reflect the shift in the relative liquidity of assets due to financial innovation.
- The certainty equivalence index is log-linear in its component assets, as is the Divisia index, but assigns a weight of unity to currency.
- The appropriate cost of using a monetary asset is its user cost.
- There are several criteria for judging among monetary aggregates. Two of the most important of these are the stability of their demand function and their performance in explaining nominal national income.
- The empirical evidence tends to favor Divisia aggregation for broader monetary definitions while favoring simple-sum aggregation for the narrower examples.
- Empirical evidence clearly shows that changes in the money stock do cause changes in nominal national income.

Review and discussion questions

1. How would you define the liquidity of an asset, measure this liquidity, and to what uses can your definition have relevance? Compare your definition with one other definition or measure of liquidity and outline its strong and weak points.
2. The definition of money is clear enough if we consider only the transactions demand for money, but the introduction of the asset demand for money makes any attempt to distinguish sharply between M1 and other monetary assets unsatisfactory. Discuss.
3. How would you test for the substitutability of different near-monies for M1? Present at least two different procedures that have been used in the literature for this purpose. How would you compare the estimates obtained from the two procedures?
4. Compare the a priori restrictions imposed on the elasticity of substitution by the following aggregation procedures:

   (i) simple-sum aggregates
   (ii) Divisia aggregates
   (iii) VES (variable elasticity of substitution) aggregates.

   Using your intuition on the likely elasticities of substitution in your country, with demand deposits as the base asset, in which aggregate would you place the following?

   (a) saving deposits in commercial banks
   (b) saving deposits in other financial institutions (name these)
   (c) money market mutual funds sold by banks to the public (if any)
   (d) money market mutual funds sold by other institutions to the public (if any)
   (e) treasury bills

(f) short-term bonds
(g) shares of corporations

5. Can evidence that the lagged values of the money stock are significant in equations forecasting nominal GDP help in establishing the claim that changes in the money stock Granger-cause changes in nominal GDP? Discuss.
6. Can evidence that the future values of the money stock are significant in equations forecasting nominal GDP provide any information on the direction of Granger causality between changes in the money stock and nominal GDP? Discuss.
7. Given two variables Y and X, specify the Granger test for the direction of causality between them. What did Sims's study on Granger causality between money and national income for the United States show?

   How would you also test for the possibility that a third variable Z also Granger-causes one or both of Y and X?
8. "Money–demand functions have been shifting in ways that make them unsuitable for selecting the appropriate monetary aggregate for a given economy." Report on some studies that have used the stability of the money–demand function for judging among monetary aggregates but have arrived at conflicting results. In view of such conflicting findings, should we abandon 'data mining' for the 'best' monetary aggregate and stick to an a priori or theoretical definition of money? Discuss.
9. Conduct the Sims–Granger test for the direction of causality between simple-sum and Divisia monetary aggregates and income in your economy and interpret its results for the direction of causality and the choice among the monetary aggregates.
10. Specify the variable elasticity of substitution (VES) function for the monetary aggregate for your economy, derive its estimation equations and estimate the VES aggregate.
11. Specify the chain-weighted Divisia function for the monetary aggregate for your economy, derive its estimation equations and estimate the Divisia aggregate.

    What tests would you use to select between your VES and Divisia estimates? Carry out at least two of them and discuss their results.

References

Anderson, L.C., and Jordan, J.L. "Monetary and fiscal actions: a test of their relative importance in economic stabilization." *Federal Reserve Bank of St Louis Review*, 1968, pp. 11–24.

Barnett, W.A. "Economic monetary aggregates: an application of index number and aggregation theory." *Journal of Econometrics*, 14, 1980, pp. 11–48.

Barnett, W.A., Offenbacher, E.K., and Spindt, P.A. "The new Divisia monetary aggregates." *Journal of Political Economy*, 92, 1984, pp. 1049–85.

Belongia, M.T. "Measurement matters: recent results from monetary economics revisited." *Journal of Political Economy*, 104, 1996, pp. 1065–83.

Belongia, M.T., and Chalfant, J.A. "The changing empirical definition of money: some estimates from a model of the demand for money substitutes." *Journal of Political Economy*, 97, 1989, pp. 387–97.

Belongia, M.T., and Chrystal, K.A. "An admissible monetary aggregate for the UK." *Review of Economics and Statistics*, 73, 1991, pp. 497–503.

Chetty, V.K. "On measuring the nearness of near-monies." *American Economic Review*, 59, 1969, pp. 270–8l.

Chrystal, K.A., and MacDonald, R. "Empirical evidence on the recent behavior and usefulness of simple-sum and weighted measures of the money stock." *Federal Reserve Bank of St Louis Review*, 76, 1994, pp. 73–109.

Donovan, D.J. "Modeling the demand for liquid assets: an application to Canada." *International Monetary Fund Staff Papers*, 25, 1978, pp. 676–704.

Feige, E.L., and Pearce, D.K. "The substitutability of money and near-monies: a survey of the time-series evidence." *Journal of Economic Literature*, 15, 1977, pp. 439–70.

Friedman, B., and Kutner, K.N. "Money, income, prices and interest rates." *American Economic Review*, 82, 1992, pp. 472–93.

Friedman, M., and Schwartz, A.T. *A Monetary History of the United States 1870–1960.* Princeton, NJ: Princeton University Press, 1963a.

Friedman, M., and Schwartz, A.T. "Money and business cycles." *Review of Economics and Statistics*, 45, 1963b, Supp., pp. 32–64. Comments, pp. 64–78.

Gebregiorgis, B.S., and Handa, J. "Monetary aggregation for a developing economy: a case study of Nigeria." *Journal of Developing Areas*, 38, 2005, pp. 119–43.

Goldfeld, S.M. "The demand for money revisited." *Brookings Papers on Economic Activity*, 3, 1973, pp. 577–638.

Judd, J.P., and Scadding, J.L. "The search for a stable money demand function: a survey of the post-1973 literature." *Journal of Economic Literature*, 20, 1982, pp. 993–1023.

Lebi, J., and Handa, J."Re-examining the choice among monetary aggregates: evidence from the Canadian economy." *ICFAI Journal of Monetary Economics*, 5, 2007, pp. 57–78.

Rotemberg, J.J. "Commentary: monetary aggregates and their uses." In M.T. Belongia, ed., *Monetary Policy on the 75th Anniversary of the Federal Reserve System*. MA: Kluwer Academic Publishers, 1991, pp. 223–31.

Rotemberg, J.J. "Monetary aggregates, monetary policy and economic activity: commentary." *Federal Reserve Bank of St Louis Review*, 75, 1993, pp. 36–41.

Rotemberg, J.J., Driscoll, J.C., and Porterba, J.M. "Money, output and prices: evidence from a new monetary aggregate." *Journal of Economic and Business Statistics,* 13, 1995, pp. 67–83.

Serletis, A., and Robb, A.L. "Divisia aggregation and substitutability among monetary assets." *Journal of Money, Credit and Banking*, 18, 1986, pp. 430–46.

Serletis, A., and Molik, T.E. "Monetary aggregates and monetary policy." In Bank of Canada, ed., *Money, Monetary Policy and Transmission Mechanisms*, 1999.

Sims, C. "Money, income and causality." *American Economic Review*, 62, 1972, pp. 540–52.

Sriram, S.S. "Survey of literature on demand for money: theoretical and empirical work with special reference to error-correction models." *International Monetary Fund Working Paper* No. 64, 1999.

Swofford, J.L., and Whitney, G.A. "Nonparametric tests of utility maximization and weak separability for consumption, leisure and money." *Review of Economics and Statistics*, 69, 1987, pp. 458–64.

Varian, H.R. "Nonparametric tests of consumer behavior." *Review of Economic Studies*, 50, 1983, pp. 99–110.

8 The demand function for money

A number of issues have to be resolved prior to the empirical estimation of the demand for money. Among these are the use and estimation of expected and permanent income and the treatment of lags in money demand. For the former, this chapter covers the use of rational expectations. Adaptive expectations are used for the measurement of permanent income. Costs of adjusting money balances lead to lags in the adjustment of actual to desired money balances. The simplest forms of lags are the first-order and second-order (linear) partial adjustment models.

This chapter also extends the money demand function to the open economy and investigates currency substitution and capital mobility.

Key concepts introduced in this chapter

- Permanent income
- Expected income
- Rational expectations
- Adaptive expectations
- General autoregressive model
- Lucas supply rule
- Keynesian supply function
- Partial adjustment models
- Autoregressive distributive lag model
- Currency substitution and capital mobility

Milton Friedman's money demand function, presented in Chapter 2, argued that permanent income is one of the determinants of the demand for money. Other studies assume that the individual's planned money balances are a function of his expected income during the period ahead. While the data on the actual past and present levels of national income is readily available, data on the expected and permanent income are not observable. This data has to be either generated or proxied in estimating the demand function for money.

Further, while the theoretical analyses of Chapters 2 to 6 provided the three basic specifications of the demand function for desired balances, there could be significant costs of reaching the desired levels in each period, so that the actual balances held may differ from

those desired. This leads to the consideration of partial adjustment and lags in the money-demand function. Since our aim is to explain the actual balances held, the differences between the desired and actual money holdings and the procedures for handling the lags that occur in this process need to be examined.[1] In recent years, these issues have been pushed aside, though not addressed, by the increasing use of cointegration and error-correction estimation techniques.

While the money demand analyses of the preceding chapters established the arguments of the money demand function, they did not specify its specific functional form. This chapter introduces three of its more commonly used basic functional forms in empirical analyses for the closed economy. It then proceeds to the money demand function for the open economy under the heading of currency substitution.

Section 8.1 starts with three basic money demand functions, with actual income, expected income and permanent income as the scale variable. For the latter, Section 8.2 presents the rational expectations hypothesis for estimating expected income. Section 8.3 presents the adaptive expectations procedure for deriving permanent income and Section 8.4 lists the regressive and extrapolative procedures. Sections 8.5 to 8.8 present the partial adjustment model and the general autoregressive model. Section 8.9 focuses on the money demand function in the open economy.

8.1 Basic functional forms of the closed-economy money demand function

Monetary theory provides the variables that determine money demand but does not specify the particular form of the money demand function. The analysis of the demand for money in the preceding chapters implied that this demand depends on an income or wealth variable, often also called the "scale variable," and on the rates of return on alternative assets. Since these rates of return are closely related to each other, so that including several of them in the same regression induces multicollinearity (discussed in the next chapter), the money demand equation that is usually estimated avoids multicollinearity by simplifying the estimating equation to include only one interest rate. With this simplification, and using actual income as the simplest form of the scale variable, the money demand function is:

$$m^{\mathrm{d}} = m^{\mathrm{d}}(y, R)$$

where:

m^{d} = demand for real balances
y = actual real income
R = nominal interest rate

There is no real theoretical basis for assuming the form of this function to be linear, log-linear or non-linear in some other way. However, for reasons of convenience in estimation, the linear and log-linear functional forms are the most commonly used ones. This section compares these functional forms and points out the differences between them. It ignores, for simplification, the possibility of lags and expectations and assumes that money demand depends only upon current income and a nominal interest rate.

1 Cuthbertson (1985, Ch. 3) is a good adjunct to this chapter for the treatment of adjustment lags and expectations.

To start, consider the following simple specific forms of the money demand function, with μ as the random term. The subscript t has been omitted as being unnecessary for the discussion.

$$M/Y = a_0 + a_R R + \mu \tag{1}$$

$$M = a + a_R R + a_y y + a_P P + \mu \tag{2}$$

$$m = a + a_R R + a_y y + \mu \tag{3}$$

(1) assumes that the elasticity of the demand for money with respect to nominal income – and hence with respect to both prices and real income – is unity. (3) assumes that this elasticity is unity with respect to the price level but not necessarily so with respect to real income. (2) does not make either assumption.

(3) is the only function consistent with the discussion in earlier chapters that the individual's demand for money balances is in real rather than in nominal terms. Proceeding further with (3), money demand in a world where commodities and money are substitutes would also depend upon the expected rate of inflation π^e, so that (3) would be modified to:

$$m = a_0 + a_R R + a_y y + a_\pi \pi^e + \mu \tag{4}$$

Other variables, such as the expected exchange rate depreciation to take account of currency substitution in the open economy, as is done later in this chapter, could be introduced in a similar manner on the right-hand side of (4).

The money-demand functions are often estimated in a log-linear form. The log-linear form corresponding to (3) would be:

$$\ln m = \ln a_0 + \alpha \ln R + \beta \ln y + \ln \mu \tag{5}$$

A variant of (5) replaces $\ln R$ by $\ln(1 + R)$ since R is usually between 0 and 1, so that its logarithmic value would be a negative number whereas $\ln(1 + R)$ would be positive. (5) is identical to:

$$m = a_0 R^\alpha y^\beta \mu \tag{6}$$

This functional form is the well-known *Cobb–Douglas* functional form. It was implied by the inventory analysis of the transactions demand for money, though not by the speculative or the precautionary demand analyses. In (5) and (6), the elasticity of the demand for real balances is α with respect to R and β with respect to y. A variant of (6) is:

$$\ln m = \ln a_0 + \alpha R + \beta \ln y + \ln \mu \tag{7}$$

(7) does not require taking the log of the interest rate since doing so would yield negative values when the values of R lie between 0 and 1. However, note that (7) translates to:

$$m = a_0 e^{\alpha R} y^\beta \mu \tag{8}$$

Since (6) and (8) are different and are unlikely to perform equally well, the researcher has to choose between them. There is no theoretical basis for doing so, with the result that the one that gets to be reported often depends upon its relative empirical performance for the data being used.

8.1.1 Scale variable in the money demand function

Current income as the scale variable

The linear form of the demand function for real balances, with current income as the scale variable, is:

$$m^{d}_{t} = a_0 + a_y y_t + a_R R_t + \mu_t \quad a_0, a_y > 0, a_R < 0 \tag{9}$$

where μ is the random disturbance. (9) would become log-linear if each of the variables and μ_t were in logs.

Expected income as the scale variable

Another money demand function that is in common usage replaces current income by expected income. A demand function with expected income as its scale argument is:

$$m^{d}_{t} = a_0 + a_y y^{e}_{t} + a_R R_t + \mu_t \quad a_0, a_y > 0, a_R < 0 \tag{10}$$

In (10), at the *beginning* of the period, m^{d}_{t} are the planned real balances for the period ahead, y^{e}_{t} is the expected income for the period. While we could have also introduced the interest rate in terms of its expected value, this is rarely done. The current practice is to estimate expected income y^{e}_{t} using the rational expectations hypothesis (REH).

Permanent income as the scale variable

As against current or expected income, Friedman's (1956) theoretical analysis of the demand for money presented in Chapter 2 implied that this demand depends upon wealth, or its proxy, permanent income, and on interest rates. For Friedman's analysis, the basic form of the demand function for real balances with permanent income is:

$$m^{d}_{t} = m^{d}(y^{p}_{t}, R_t)$$

where y^{p}_{t} is permanent income, which can be interpreted as the average expected income over the future. The simplified linear (or log-linear) form of this demand function for real balances is:

$$m^{d}_{t} = a_0 + a_y y^{p}_{t} + a_R R_t + \mu_t \quad a_0, a_y > 0, a_r < 0 \tag{11}$$

Since data on the observed values of y^{p}_{t} does not normally exist, Friedman used the adaptive expectations hypothesis for deriving permanent income. Though the REH can be used as an alternative procedure for doing so, adaptive expectations seem more appropriate for estimating permanent income since the latter is best interpreted as the *average* expected value of income, rather than merely as expected income for the period ahead. Correspondingly, m^{d}_{t} in (11) should be interpreted as the *average* expected amount of desired real balances. The adaptive expectations procedure for constructing permanent income is explained in Section 8.3.

Note that the three scale variables in (9) to (11) are different, so that their estimation will yield different coefficients. Further, even their stability properties may differ. As discussed

later in this chapter and in Chapter 9, the time series for several variables, including money and income, tend not to be stationary. The appropriate technique for such variables is cointegration analysis, which is a maximum likelihood vector autoregressive (VAR) technique. Such estimation ignores altogether the distinction between expected and permanent income, so that the prior application of the rational and adaptive expectations procedures is not needed if the money demand function is estimated using cointegration techniques. However, its accompanying error-correction estimation does incorporate adjustment lags and adaptive expectations.

8.2 Rational expectations

8.2.1 Theory of rational expectations

The rational expectations hypothesis (REH), first proposed by Muth (1961), is stated in various forms. One way of stating it is that the individual uses all the available information at *his* disposal in forming his expectations on the future values of a variable. Since individuals often have to – or choose to – operate with very limited information, the relevant information set is sometimes specified to be one of maximizing profit. In any case, the available information set is assumed to include the knowledge of the *relevant theory*,[2] with the rationally expected value of the variable being its value as predicted by this theory. The REH asserts that deviations of the actual from the theoretically predicted value will be randomly distributed with a zero mean and be uncorrelated with the available information and with the theoretically predicted value.

Note that the relevant theory will commonly determine the non-random prediction of a variable as a function of the parameters, the past values of the endogenous variables and the past, current and future values of the exogenous variables. Of these, the future values of the exogenous variables will usually not be known to the individual and their rational expectations values will be needed, so that the relevant theory for them will also have to be specified. In practical terms, the REH can be restated as: the expected values of the endogenous variables will be those predicted by the relevant theory, given the data on the past values of the endogenous variables, those on the past and current values of the relevant exogenous variables, and the rationally expected future values of the relevant exogenous variables.

Designate the rationally expected value of y^{e}_{t} predicted by the relevant theory as y^{T}_{t}, where the superscript T stands for the relevant theory. Since y^{T}_{t} takes account of all the information available to the individual, the REH asserts that the deviation of the actual value y_t from y^{T}_{t} will be random with a zero expected value and will be uncorrelated with the available information and, therefore, with y^{T}_{t} which is based on that information. The following incorporates the above statements in a set of simple equations to show

2 The switch from "information available to the individual" to "knowledge of the relevant theory" is a massive leap. The former implies a "subjective theory," which is likely to differ from those held by others, whereas the latter relates to an "objective theory" common to all individuals and based on accurate knowledge.

Note also that the word "relevant" is a loaded one. It means the *correct* theory for the economy or market in question. However, the correct theory is hardly ever known, as witnessed by the differing macroeconomic schools on the determination of national income and inflation or by disagreements even among the exponents of a given school on the actual values of the structural and reduced form coefficients of the model.

the various assumptions and steps in deriving the rationally expected value y^{e*}_t of the variable y_t.

Since the rationally expected value y^{e*}_t – with ^{e*} standing for the rationally expected value – is assumed to be determined by the value y^{T}_t predicted by the relevant theory T, we have:

$$y^{e*}_t = y^{T}_t \tag{12}$$

Since the REH assumes that the actual value y_t differs from the prediction of the relevant theory T by an error that is random and not correlated with any available information, we have:

$$y_t = y^{T}_t + \eta_t \tag{13}$$

where:

$$E\eta_t = 0 \tag{14}$$

$$\rho(y^{T}_t, \eta_t) = 0 \tag{15}$$

y_t = actual income
y^{e}_t = expected income
y^{e*}_t = rationally expected value of income
y^{T}_t = expected income predicted by the relevant theory
$E\eta_t$ = mathematical expectation of η_t
$\rho(y^{T}_t, \eta_t)$ = correlation coefficient between y^{T}_t and η_t.

(12) and (13) imply that:

$$y_t = y^{e*}_t + \eta_t \tag{16}$$

Taking the mathematical expectation of (16), with $E\eta_t = 0$ from (14), and using (12) gives:

$$Ey^{e*}_t = Ey_t = y^{T}_t \tag{17}$$

If y^{e*}_t and y^{T} are assumed, as is often done, to be single valued, (17) becomes:

$$y^{e*}_t = Ey_t = y^{T}_t \tag{18}$$[3]

To implement (18) empirically, the rationally expected value y^{e*}_t can be obtained by estimating y_t using the function implied by the theory for its determination, and taking its expected value Ey_t.[4] This procedure is illustrated below and will also be applied in Chapter 17 in a macroeconomics context.

3 This is called the *weak* version of the REH. A *strong* version of it is that $F(y^{e*}_t) = f(y_t)$, where F and f specify the respective frequency distributions. This version requires that the distribution of expected income is the same as that of actual income except for a random term.

4 In purely theoretical analysis, y^{T} is used for deriving y^{e*} if y^{T} is single valued. If it is not, y^{e*} is replaced by Ey^{T}.

The "relevant theory"

A fundamental question in applying the REH is about the definition of the term "relevant theory." To an economist who believes that the economy tends to be at full employment, even though it is currently not in that state, the relevant theory for forming expectations on aggregate output is that the economy will be at the full-employment level. Consequently, the full-employment output will be the rationally expected one, so that the appropriate procedure would be to solve the model or theory for its full-employment state and substitute it for the expected output or real income. This procedure is the one adopted by economists in the modern classical approach.

However, for economists who believe that the economy is rarely, if ever, exactly in full employment, the rational expectation of next period's real income will not be one of full employment. The theory needed for their rational expectations of income would be a theory of the non-random part of the expected level of actual income, since this is the level that would differ from next period's actual income by a random term. Keynesian economists follow this line of thinking and need to specify a theory of the expected value of actual output for the period in question.

Hence, the application of the REH will yield different values of rationally expected output, depending upon the underlying assumption of the continuous existence or frequent absence of full employment. While the REH at the conceptual level can be and is used by both classical and Keynesian economics, its application, even in the context of an otherwise identical model (e.g. the IS–LM one), provides different predictions of the expected future income for the two approaches.

To proceed further, (18) can be used to construct the estimate of y^{e*}_t by using the relevant theory to specify the determination of y^{T}_t. We illustrate this use of the theory by incorporating two alternative theories on the relationship between output and the rate of increase in the money supply. The first theory will be the Lucas supply rule, which underpins modern classical macroeconomics, and the second one will be the Keynesian theory.[5]

8.2.2 Information requirements of rational expectations: an aside

There is considerable dispute in the literature about the information requirements for rational expectations. The information available to any given individual varies considerably, *inter alia,* with the individual's level of education and interest, the openness of the society and the operating technology of information, as well as the losses from basing actions on inadequate, vague and inaccurate information. The actual amount of information at the disposal of the individual can vary from almost non-existent hard information[6] to extensive knowledge. The REH is meant to apply to all cases, regardless of the extent and accuracy of the available information.

Skeptics about the REH have argued that it requires that the possible future outcomes are well anticipated and that economic agents are assumed to be superior economists and

5 See Chapters 15 and 17 for these models and rules.

6 For instance, in many underdeveloped or developing countries there is hardly any reliable information published on national income and the rate of inflation. Even if it is published, most of the inhabitants in the rural and even the urban areas may never receive or bother to acquire this information.

statisticians, capable of analyzing the future general equilibrium of the economy (Arrow, 1978). However, the supporters of rational expectations reject such criticism and claim that:

> The implication that economic agents or economists are omniscient cannot fairly be drawn from Muth's profound insights. … Rational expectations are profit maximizing expectations. … If the past proves to be a very imperfect guide to the future, then theory and practice will be inaccurate.
>
> (Kantor, 1979, p. 1424).

> It is, however, incorrect to assume that rational expectations regards errors as insignificant or absent. The implication of rational expectations is that the forecast errors are not correlated with anything that could profitably be known when the forecast is made.
>
> (Kantor, 1979, p.1432).

Another view of rational expectations is provided by Robert Lucas, who popularized its usage in macroeconomics. The *Economist* website reported that Lucas at one time said that:

> [Rational expectations] doesn't describe the actual process people use trying to figure out the future. Our behavior is adaptive. We try some mode of behavior, if it is successful, we do it again. If not, we try something else. Rational expectations describe the situation when you've got it right.[7]

This interpretation means that, for most of the time spent in figuring out the future and acting on one's expectations, we would not have rational expectations with its critical property that the errors between the actual and the expected value would be random. Since rational expectations will only hold eventually ("when you have got it right"), they should be restricted only to the long-run analysis. They will not apply over short periods and in the short run. This interpretation of rational expectations is not consistent with the macroeconomic literature on it, including Lucas's own contributions on the short-run macroeconomic model. It is also not consistent with the use and analyses of the Lucas supply rule presented in this chapter and in Chapter 14. We shall henceforth ignore this interpretation.

Assessing the validity of the rational expectations hypothesis

The insight behind rational expectations at its conceptual level – that is, when an individual's expectations are based on all his available information – is undeniable. However, part of the available information comes from our understanding of the past and the present, which is itself incomplete and imperfect, as witnessed by the prevalence, even in hindsight, of different theories to explain any given observation. In addition, knowledge of the future is even more uncertain; as the quote at the end of this subsection argues, for the future, we don't even know what we don't know. Given the increasing increment of this degree of ignorance for periods further ahead, short-term (i.e. for the next quarter or so) predictions tend to do better than for periods further ahead. But, for these, the persistence forecast (i.e. the immediate future will

7 http://economistsview.typepad.com/economistsview/2007/09/expect-the-unexp.html, downloaded and printed 12 September 2007.

be like the immediate past except for random variations) does quite well – and usually better than predictions based on any theory.

In the rational expectations hypothesis, the leap from the "subjective/personal theory" based on the available information to the assumption of the "relevant theory," common to all individuals as well as being the accurate theory, is a massive one. This assumption is also likely to be invalid. The exponents of the REH, with Kantor among them, focus on the former, whereas its critics, with Arrow among them, focus on the latter. Leaving aside the doctrinal disputes, the *empirical issue* boils down to a question of the usefulness or profitability of *acting* on one's rational expectations. This usefulness can be extremely limited when – without knowledge of the relevant theory and without good reliable information on the past values of the endogenous and exogenous variables, or on the relevant future values of the exogenous variables – the known paucity of information indicates that the actual error in the rationally expected value of a variable can be large[8] relative to the mean expected value of the variable, so that acting on the basis of the rationally expected value of the variable may not turn out to be a prudent exercise.[9] Conversely, if the information available is quite complete and the subjective probabilities are known to approximate the objective ones, the rational expectations could be an appropriate basis for action.

An interesting take on the nature of uncertainty and how it limits the reliability and usefulness of rational expectations for decisions is provided by the following quote:

> There are *things that we know*. There are [also] *known unknowns*; that is to say that there are things we now know that we don't know. But there are also *unknown unknowns* – things that we do not know we don't know. So when we do the best we can and we pull all the information together, and we then say, "Well, that is basically what we see as the situation," that is really only the known knowns and the known unknowns. And [as time passes] we discover a few more of those unknown unknowns. There is another way to phrase that, and that is that the absence of evidence is not evidence of absence.
>
> (Donald Rumsfeld in a news conference, June 2002; italics added).

8.2.3 Using the REH and the Lucas supply rule for predicting expected income

This rule assumes the modern classical model, with the labor market being in long-run equilibrium at full employment and with deviations in real national output from its full employment level y^f occurring only due to errors in predicting the actual level of the

8 This need not mean that their mean value is not zero. It is that they can take relatively large absolute values.

9 Acting on the basis of the same rationally expected value of a variable may be very prudent behavior when the information on which it is based is fairly complete and reliable. Acting on it may be foolhardy behavior when the information is scanty and represents a shot in the dark rather than an adequate basis for action.

A distinction is drawn here between "prudence" and "profit maximizing" in the presence of an acute lack of information; I may know nothing about horse racing and about the horses entered in a race, yet I can on the basis of the available information – say on the pleasing and non-pleasing colors of horses – specify my subjective probabilities of the performance of the horses in the race and place my profit-maximizing bets on the basis of these probabilities. Prudence may instead dictate that I recognize the vagueness and inadequacy of my information underlying my probabilities and that I do not bet at all.

Another distinction is drawn between *holding* rational expectations and *acting* on those expectations. Rational expectations can always be held, as long as the required probabilities are subjective and not objective ones. However, experience may have taught the individual not to act on them, or to act on them after due allowance and caution for a large margin of uncertainty, ambiguity and error.

money supply. One form of the Lucas supply rule[10] specifies the relevant theory for the determination of output y in period t as:

$$y^{\mathrm{T}}{}_t = y^{\mathrm{f}}_t + \gamma(M_t - M^{\mathrm{e}}{}_t) \tag{19}$$

where:

y^{f}_t = full employment level of output in t
M_t = nominal money stock in t
$M^{\mathrm{e}}{}_t$ = expected value of the nominal money stock in t
$M^{\mathrm{e}*}{}_t$ = rational expectations of M_t, formed in $t-1$

so that the rational expectation of income, with the Lucas supply rule as the relevant theory for its determination, is:

$$y^{\mathrm{e}*}{}_t = y^{\mathrm{f}}_t + \gamma(M_t - M^{\mathrm{e}*}{}_t) \tag{20}$$

Use of (20) for predicting rationally expected income requires using the relevant theory to determine $M^{\mathrm{e}*}{}_t$. The relevant theory depends upon the monetary policy being pursued by the monetary authority.[11] In the context of an exogenous money supply, the central bank controls the money supply and can determine the money supply in the economy on the basis of a "rule" or function. Assume this to be the case, and that the relevant theory for the central bank's money supply rule M^{T} is:

$$M^{\mathrm{T}}{}_t = \Psi_0 + \Psi_1 u_{t-1} + \Psi_2 M_{t-1} \tag{21}$$

where u_t is the unemployment rate (or the output gap between full employment and actual output) in period t. Designating the random error in M_t as ξ_t, (21) leads to the specification of M_t as:

$$M_t = \Psi_0 + \Psi_1 u_{t-1} + \Psi_2 M_{t-1} + \xi_t \tag{22}$$

Estimating (22) will provide the estimated values of the coefficients $\Psi_i, i = 0, 1, 2$, as $\hat{\Psi}$. These estimated coefficients can be used to estimate EM_t, which yields the rationally expected value $M^{\mathrm{e}*}{}_t$, as:

$$\begin{aligned} \hat{M}^{\mathrm{e}*}{}_t &= E\hat{M}_t \\ &= \hat{\Psi}_0 + \hat{\Psi}_1 u_{t-1} + \hat{\Psi}_2 M_{t-1} \end{aligned} \tag{23}$$

Since $M_t - M^{\mathrm{e}*}{}_t = M_t - \hat{\Psi}_0 + \hat{\Psi}_1 u_{t-1} + \hat{\Psi}_2 M_{t-1} = \hat{\xi}_t$, where $\hat{\xi}_t$ is the estimated value of ξ_t, (19) implies that:

$$y_t = y^{\mathrm{f}}_t + \gamma\hat{\xi}_t + \eta_t \tag{24}$$[12]

10 The macroeconomics of the Lucas supply rule are covered in Chapter 14. This rule makes the output gap $(y_t - y^{\mathrm{f}}_t)$ a positive function of the deviation of the price level from its expected value. Adding the assumption that the price level is a function of the money supply leads to the specification of the Lucas supply rule assumed in the text.

Note that the Lucas supply rule is not consistent with the various specifications of the Keynesian model.

11 See Chapters 10, 11, 13 and 17 for the money supply theories.

12 $\hat{\xi}_t$ is the estimated value of the unanticipated money supply obtained from the estimation of (22).

In the estimation of (24), y^{f}_{t} is replaced by a constant term. The estimation of (24) then yields the estimated values $\hat{y}^{f}_{t}$ and $\hat{\gamma}$, so that the rationally estimated value y^{e*}_{t} of y^{e}_{t} can now be derived from:

$$\hat{y}^{e*}_{t} = \hat{y}^{f}_{t} + \hat{\gamma}\,\hat{\xi}_{t} \qquad (25)^{13}$$

The procedure for the estimation of the money demand function using the REH and the Lucas supply function

In the above illustration of the REH, it was necessary to estimate the money supply function (22) in order to estimate the error in the expected value of the unanticipated money supply; then to use this value in (24) to estimate the expected value of real output/income; this was followed by the use of this estimated value of real income in the regression for the money demand function (10). Hence, estimating the money-demand function – using the REH and the Lucas supply rule – required estimation of at least three equations in a stepwise procedure. The reliability of its estimates of the money demand coefficients in (10) would therefore depend upon the proper specification of the model for y^{T}_{t} and of its subsidiary equations for the money supply function, as well as of the reliability of the data and the estimating techniques used at the various stages. Clearly, there is considerable scope for possible errors in specification and biased estimation.

Keynesians believe that one of the sources of errors in the above estimation is the specification of the Lucas supply rule as the "relevant theory" for the determination of income. They believe that a Keynesian supply function is the appropriate theory. The following presents their approach.

8.2.4 Using the REH and a Keynesian supply function for predicting expected income

A simple form of the Keynesian supply rule[14] for the context of an exogenous money supply is:

$$y^{T}_{t} = y_{t-1} + \beta(Dy_{t-1}) \cdot M_{t} \quad \beta \geq 0 \qquad (26)$$

where $Dy_{t-1} = (y^{f}_{t} - y_{t-1})$. (26) specifies that real income/output depends upon the actual money supply, rather than only on the unanticipated change in the money supply. Further, this impact depends on the prior state of the economy, with this state captured by the lagged output gap[15] Dy_{t-1}. If the prior state is one of full employment, $Dy_{t-1} = 0$, so that changes in the money supply will not change output. The larger the output gap, the larger the impact of the money supply on output.

13 Note that this procedure allows for the possibility that the mean value of the *estimated* errors in the money supply equation is not zero. However, on an a priori basis, it sets the expected value of the random error term in the output equation at zero. There is admittedly an inconsistency in this treatment of the errors between the two equations. The alternative would be to set, on an a priori basis, the expected values of both terms at zero. But then, (25) would become $y^{e*}_{t} = \hat{y}^{f}_{t}$, which poses other problems for the preceding analysis.

14 See Chapters 15 and 17 for discussion and further use of this concept.

15 Alternatively, the output gap can be replaced by the deviation of the actual unemployment rate from the natural one.

The stochastic form of this relationship is:

$$y_t = y_{t-1} + \beta(Dy_{t-1})M_t + \eta_t \tag{27}$$

In estimation, (27) uses the *actual* value of M_t as a regressor and therefore does not require the prior estimation of the coefficients of the money supply function or of the anticipated money supply.[16] Consequently, (27) requires the estimation of only one equation for estimating the expected value of income, rather than estimation of two under the Lucas supply rule. Again assuming the REH, using the estimated value of β from (27) provides the estimate of the (Keynesian) rationally expected value $y^e{}_t$ as:

$$\hat{y}^{e*}{}_t = \hat{y}_{t-1} + \hat{\beta}M_t \tag{28}$$

Compare (25) and (28). Note that both provide the "rationally expected values of income" but under different theories as being the "relevant" one.

The procedure for the estimation of the money demand function using the REH and the Keynesian supply function

Proceeding further with (28), the rationally expected value of y_t can now be inserted into the money demand function (10) to estimate the latter. Hence, the use of the Keynesian supply function and the REH requires only a two-step procedure for estimation of money demand.

8.2.5 Rational expectations – problems and approximations

While rational expectations require that $y^e{}_t$ be based on all available information, the information available to the economist is different from that available to the individual. Further, the economist generally deals with aggregates – for example, with aggregate money demand or national income – rather than with the money demand or income of any given individual, so that what the relevant information set should be is not always clear. Furthermore, there are disputes among economists as to the relevant theory, or at least to the theory held by the public.[17] Even when there is agreement among economists – admittedly a rare occurrence – on the general form of the theory, there is usually disagreement on the values of the coefficients of the model *and* on the *expected* values of the exogenous variables for the period ahead. Even the data on the lagged values of the endogenous variables is usually approximate and subject to revision, sometimes for several years after the data period. These problems and disputes render rational expectations a blunt procedure at the estimation level, and its applications subject to doubt and disputes.

16 A problem arises if M_t is not known at the beginning of t and has to be replaced by its estimate. This estimate would require knowledge of both the anticipated and unanticipated elements of money supply.

17 Evidence of this comes from disputes between the classical and the Keynesian paradigms, and within each of these paradigms. The periodic switches that occurred between them in the 1930s and 1940s from the classical to the Keynesian paradigm, and then from the Keynesian to the classical one in the 1970s and 1980s, is historic testimony to the fact that economists do not know the true model of the economy – and therefore do not know the relevant model for determining national income or the rate of inflation, or other macroeconomic variables. If the economists do not know the true model, the public can hardly be expected to form expectations on the basis of the "relevant" model of the economy.

Further, economists as a group are notorious for their forecasting failures.

In view of the absence of direct quantitative data on expected income and problems with applying rational expectations at the empirical level, some researchers choose to proxy y^e_t in various ways. Two examples of this are:

1 Use the actual income y_t as a proxy for y^e_t, since the two differ only by a random term whose expected value is zero under rational expectations.
2 Use the autoregressive model:

$$y_t = \delta_0 + \delta_1 y_{t-1} + \delta_2 y_{t-2} + \cdots + \mu_t \tag{29}$$

and then use $y^e_t = Ey_t$ and the estimated coefficients of (29) to estimate y^e_t. The justification for this approximation to rational expectations is that the past experience of income itself is likely to be the dominant part of the relevant information set of the individual and the public, and the past values of income are likely to be most important determinant of current income in the relevant model.

While the REH at the conceptual level is very appealing, such approximations in empirical applications do reduce its distinctiveness from the rivals to the REH and are not recommended – unless there is no better choice.

8.3 Adaptive expectations for the derivation of permanent income and estimation of money demand

The specification of permanent income

In order to illustrate the application of adaptive expectations in money demand estimation, we shall use permanent income as the income variable in the money demand function. This function is:

$$m^d_t = a_0 + a_y y^p_t + a_r R_t + \mu_t \quad a_0, a_y > 0, a_r < 0 \tag{11}$$

The general adaptive expectations model assumes that the individual bases his permanent income on his experience of current and past actual income, so that the general function for permanent income y^p_t would be:

$$y^p_t = f(y_t, y_{t-1}, y_{t-2}, \ldots) \tag{30}$$

A simple form of (30), which has proved to be convenient for manipulation and was used by Friedman for deriving permanent income in his empirical work on consumption and money demand, is the *adaptive expectations (geometric distributed lag) function*. It specifies the functional form of y^p_t as:

$$y^p_t = \theta y_t + \theta(1-\theta) y_{t-1} + \theta(1-\theta)^2 y_{t-2} + \cdots \tag{31}$$

where $0 \leq \theta \leq 1$. Permanent income is thus specified as a weighted average of current and past incomes, with higher weights attached to the more recent incomes. Note that if $\theta = 0.40$, a weight often cited as approximating reality for annual consumption data, the weights decline in the pattern 0.4, 0.24, 0.144, 0.0864, …, so that income more than four years earlier can be effectively ignored. If actual income becomes constant, permanent income will come to equal this constant level of actual income.

Koyck transformation of the geometric distributed lag function

Lag (31) one period and multiply each term in it by $(1-\theta)$. This gives:

$$(1-\theta)y^{p}_{t-1} = \theta(1-\theta)y_{t-1} + \theta(1-\theta)^2 y_{t-2} + \theta(1-\theta)^3 y_{t-3} + \cdots \tag{32}$$

Subtracting (32) from (31) gives the equation:

$$y^{p}_{t} = \theta y_t + (1-\theta)y^{p}_{t-1} \tag{33}$$

(33) is known as the *Koyck transformation*. This transformation allows permanent income to be stated in terms of the revision of its value last period in the light of current income.

Deriving the estimation form of the money demand function

Substituting y^{p}_{t} from (33) in the money demand function (11) gives:

$$m^{d}_{t} = a_0 + a_y\theta y_t + a_y(1-\theta)y^{p}_{t-1} + a_R R_t + \mu_t \tag{34}$$

Lag each term in (34) by one period and multiply it by $(1-\theta)$. This gives:

$$(1-\theta)m^{d}_{t-1} = (1-\theta)a_0 + a_y(1-\theta)y^{p}_{t-1} + a_R(1-\theta)R_{t-1} + (1-\theta)\mu_{t-1} \tag{35}$$

Subtracting (35) from (34) to eliminate y^{p}_{t-1} gives:

$$m^{d}_{t} = a_0\theta + a_y\theta y_t + a_R R_t - a_R(1-\theta)R_{t-1} + (1-\theta)m^{d}_{t-1} + \{\mu_t - (1-\theta)\mu_{t-1}\} \tag{36}$$

where $a_y, a_R > 0$, and $0 \leq \theta \leq 1$. The objective in carrying out the above steps was to eliminate the variable y^{p} on which data is not available. (36) achieves this objective.

The estimating form of (36) is:

$$m^{d}_{t} = \alpha_0 + \alpha_1 y_t + \alpha_2 R_t + \alpha_3 R_{t-1} + \alpha_4 m^{d}_{t-1} + \eta_t \tag{37}$$

where:

$\alpha_0 = a_0\theta$
$\alpha_1 = a_y\theta$
$\alpha_2 = a_R$
$\alpha_3 = -a_R(1-\theta)$
$\alpha_4 = (1-\theta)$
$\eta_t = \{\mu_t - (1-\theta)\mu_{t-1}\}$

Note that (37) involves lagged terms in m and in R, but not in y. Further, the disturbance term in (37) is $\{\mu_t + (1-\theta)\mu_{t-1}\}$, which is a *moving average error*.

Adaptive expectations as the error-learning model

The adaptive expectations procedure in the form given by (33) can also be stated in a form known as the *error-learning model*. This form is:

$$(y^{p}_{t} - y^{p}_{t-1}) = \theta(y_t - y^{p}_{t-1}) \tag{38}$$

which specifies the *revision in permanent income* on the basis of the experienced difference or "error" between the actual income in t and the permanent income for period $(t-1)$. From (38), if $\theta = 0$, the estimate of permanent income is never revised on the basis of the experience of current income.

Assessing the relevance and validity of the adaptive expectations procedure

If we compare the rational and the adaptive expectations procedures for estimating the money-demand function, the former requires the estimation of at least two (possibly three, as in our illustration above) equations for the Lucas supply rule. However, doing so has the advantage that it allows better identification of the sources of shifts. Conversely, the adaptive expectations procedure has the disadvantage that if the parameters of the estimated money demand function shift, it is not clear whether the parameters of the money demand function or of the expected income equation have shifted. Further, in cases of monotonically increasing (decreasing) income paths, adaptive expectations induce persistent and increasing negative (positive) errors (i.e. $y_t - y^{\mathrm{p}}{}_t$) in expected income relative to actual income, so that rational individuals will revise their procedure for forming expectations away from adaptive expectations. Adaptive expectations also fail to take account of any information available to the individual about future changes in income, and are said to be (only) *backward looking*.

However, note that the adaptive expectations model, in spite of its name, really provides an estimate of the average level of future income – rather than the expected value of income for the period ahead – through its geometric distributed lag procedure, while the REH procedure provides a more appropriate estimate of the latter. The two procedures therefore provide proxies for different concepts of income, so that the choice among them should depend on the income variable, which is the appropriate scale variable in the money demand function. If the non-stochastic component of income is fluctuating and the appropriate scale variable is permanent or *average expected* income $y^{\mathrm{p}}{}_t$, the geometric distributed lag would be a better representation of this average than the rationally expected value of current income $y^{\mathrm{e}*}{}_t$.[18]

8.4 Regressive and extrapolative expectations

An alternative to adaptive expectations is the *regressive expectations* model, which specifies that:

$$y^{\mathrm{e}}{}_t = y_{t-1} + \delta(y^{\mathrm{LR}} - y_{t-1}) \tag{39}$$

where y^{LR} is the long-run level of income. Here, the expectation is that income will tend towards its long-run value.[19]

Another model of expectations is the *extrapolative expectations* one. It is that:

$$y^{\mathrm{e}}{}_t - y_{t-1} = \delta(y_{t-1} - y_{t-2}) \tag{40}$$

18 However, the rationally expected value of a permanent measure of future incomes would be better still.

19 Keynes specified such a process for interest rate expectations in the speculative demand for money: the individual expects the interest rate to move towards its long-run value.

This model assumes that income is expected to change as a proportion of the change in income in the preceding period. That is, recent *changes* – or the factors producing those changes – are expected to determine the pattern of future changes.

Whether the adaptive, regressive, extrapolative or rational expectations procedures are more appropriate depends upon how the individual forms his expectations. The adaptive expectations model seems to be the most common one for modeling permanent income, i.e. average expected income, while the rational expectations procedure is the most common one for modeling expected income over the period ahead.

8.5 Lags in adjustment and the costs of changing money balances

Lags often occur in the adjustment of money demand to its desired long-run value. There can be several reasons for such lags. Among these are: (i) habit persistence and inertia, (ii) slow adjustment of money balances due to uncertainty on whether the changes in the determinants (income and interest rates) of money demand are transitory or longer lasting, and (iii) adjustment costs, which can be monetary or non-monetary. This section focuses on adjustment costs and presents the derivation of adjustment patterns from adjustment cost functions.

First-order partial adjustment model

One reason for an adjustment lag can be the short-run cost of changing money balances. To investigate the relationship between such costs and the adjustment lag in money balances, let the individual's desired real balances be m^*_t and assume that the individual faces various types of costs of adjusting instantaneously to m^*_t. Examples of such costs are:

(i) The cost of being below or above m^*_t. For example, having inadequate balances can prevent one from carrying out purchases which require immediate payments in money.
(ii) The cost of changing actual balances from m_{t-1} to m_t.

These costs can take various forms. A simple form of these occurs when (i) has the proportional quadratic form $a(m_t - m^*_t)^2$ and (ii) has the proportional quadratic form $b(m_t - m_{t-1})^2$. Assuming these to be so, the total adjustment cost c of reaching the desired balance in period t is given by:

$$c_t = a(m_t - m^*_t)^2 + b(m_t - m_{t-1})^2 \quad a, b \geq 0 \tag{41}$$

The individual is taken to minimize this cost. The first-order condition for maximization is that:

$$\partial c_t / \partial m_t = 2a(m_t - m^*_t) + 2b(m_t - m_{t-1}) = 0 \tag{42}$$

which yields the actual balances m_t as:

$$m_t = \gamma m^*_t + (1 - \gamma) m_{t-1} \tag{43}$$

where $\gamma = a/(a + b)$. (43) can be restated in a more intuitive form as:

$$m_t - m_{t-1} = \gamma(m^*_t - m_{t-1}) \quad 0 \leq \gamma \leq 1 \tag{44}$$

(43) and (44) constitute the *first-order* (i.e. with a one-period lag only) *partial adjustment model* (*PAM*): the adjustment of real balances in period t is partial, linear and involves a one-period lag. This model suffers from the disadvantage that if m^*_t has a positive or negative trend, the divergence of actual balances from the desired ones will increase over time. Individuals would find it profitable to avoid this by abandoning the first-order PAM and using some other adjustment mechanism.[20] Therefore, the first-order PAM is inappropriate when the desired or actual balances have a strong trend component.[21] A higher-level PAM would be more appropriate in such a case.[22]

Second-order partial adjustment model

Higher-order partial adjustment models result from more complicated specifications of the adjustment costs. The *second-order* (i.e. with a two-period) *partial adjustment model* is given by the adjustment cost function:

$$c_t = a(m_t - m^*_t)^2 + b(m_t - m_{t-1})^2 + k(\Delta m_t - \Delta m_{t-1})^2 \quad a, b, k > 0 \tag{45}$$

$$= a(m_t - m^*_t)^2 + b(m_t - m_{t-1})^2 + k(m_t - 2m_{t-1} + m_{t-2})^2 \tag{46}$$

where $\Delta m_t = m_t - m_{t-1}$, and $k(\Delta m_t - \Delta m_{t-1})^2$ is additional to the adjustment costs (i) and (ii) and represents the cost of continual adjustments over time in balances. Minimizing (46) – that is, setting the partial derivative of c with respect to m equal to zero – and solving implies that:

$$m_t = \gamma_1 m^*_t + \gamma_2 m_{t-1} + (1 - \gamma_1 - \gamma_2) m_{t-2} \tag{47}$$

where:

$\gamma_1 = a/(a + b + k)$
$\gamma_2 = (b + 2k)/(a + b + k)$

Since (47) has a two-period lag, it provides the second-order partial adjustment model.

Error feedback model

A further elaboration of these models is obtained if, in addition to the earlier types of costs, the costs of continual adjustment were less when the actual changes Δm_t were in the same direction as the desired changes Δm^*_t. A specification of such a cost function would be:

$$c_t = a(m_t - m^*_t)^2 + b(m_t - m_{t-1})^2 - k\Delta m^*_t(m_t - m_{t-1}) \quad a, b, k > 0 \tag{48}$$

20 From a rational expectations perspective, partial adjustment models are inappropriate since they are backward looking and do not take account of the available information on the expected future changes in desired balances. The individual may find it profitable to make changes in current balances to provide for future changes in desired balances.

21 The first-order PAM imposes the same response pattern to the changes in desired balances, irrespective of whether such a change is due to changes in income or in interest rates. In fact, the adjustment costs may differ depending upon the source of the changes in the desired balances.

22 This is desirable to detrending the data and using the first-order PAM.

In this case, the demand for actual balances would be:

$$m_t = m_{t-1} + \gamma_1(m^*_t - m_{t-1}) + \gamma_2(m^*_t - m^*_{t-1}) \tag{49}$$

where $\gamma_1 = [a/(a+b)]$ and $\gamma_2 = [k/\{2(a+b)\}]$. (49) is another form of PAM and is the *error feedback model*.

Assessing the validity of partial adjustment models

The various types of adjustment cost functions depend upon the notion that it is costly for the individual to change money balances. As in the inventory model of transactions balances in Chapter 4, this cost is the sum of both monetary and non-monetary costs and will differ for the different definitions of money. In practice, in modern financially developed economies with internet banking, the costs of converting to M1 from savings deposits and other near-money assets have become virtually zero, so that there should not be any significant adjustment lags at the level of the individual. The costs of changing M2 can be similarly very small and may be of little consequence at the individual's level. This is especially so when such costs are compared with those of changing the individual's stock of commodities or labor supplied.[23] The costs of adjustment for monetary aggregates usually become significant only when such adjustment involves converting bonds or commodities into a monetary asset. But this happens rarely for meeting desired changes in the demand for M1 and M2, so that the practice of using PAM models, especially for the narrower definition of money, may be of questionable value. However, the lagged adjustment implied by these models is part of the error-correction modeling within the currently popular cointegration technique.

8.6 Money demand with the first-order PAM

If there exist adjustment costs in changing money holdings, these costs should properly be incorporated into the structure of the individual's decision processes and the demand for money holdings be derived after such incorporation. However, this can prove to be analytically intractable, so that the usual procedure is to derive the demand function separately from the adjustment function and then combine them. This is the procedure followed here.

Assume that the individual's demand for real balances depends on current real income y and the nominal interest rate R, so that the demand function is:

$$m^*_t = a_0 + a_y y_t + a_R R_t + \mu_t \quad a_0, a_y > 0, a_R < 0 \tag{50}$$

where μ is white noise. m^*_t are the desired real balances in the absence of adjustment costs.

Further, assume the first-order PAM, which is:

$$m_t - m_{t-1} = \gamma(m^*_t - m_{t-1}) \quad 0 \leq \gamma \leq 1 \tag{44}$$

Substituting (44) into (50) to eliminate m^*_t gives:

$$m_t = a_0\gamma + a_y\gamma y_t + a_R\gamma R_t + (1-\gamma)m_{t-1} + \gamma\mu_t \tag{51}$$

23 This idea was explored in Chapter 6 in the context of buffer stock models of money.

where a_0, $a_y > 0$, $a_R < 0$ and $0 \leq \gamma \leq 1$. The estimating form of (51) is:

$$m^{\mathrm{d}}{}_t = \beta_0 + \beta_1 y_t + \beta_2 R_t + \beta_3 m_{t-1} + \xi_t \tag{52}$$

where $\beta_0 = a_0\gamma$, $\beta_1 = a_y\gamma$, $\beta_2 = a_R\gamma$, $\beta_3 = (1-\gamma)$ and $\xi_t = \gamma\mu_t$.

The estimating equations (37) and (52) should be compared to see the effects of adaptive expectations versus those of the first-order PAM on the estimated money demand equation. Each introduces the lagged money balances m_{t-1} into this equation, but adaptive expectations also introduce the lagged interest rate R_{t-1}. The disturbance terms also have different properties.

8.7 Money demand with the first-order PAM and adaptive expectations of permanent income

Assume now that money demand depends on permanent income, so that our model consists of the following three equations:

$$m^*{}_t = a_0 + a_y y^{\mathrm{p}}{}_t + a_R R_t + \mu_t \quad a_0, a_y > 0, a_R < 0 \tag{11}$$

$$y^{\mathrm{p}}{}_t = \theta y_t + (1-\theta) y^{\mathrm{p}}{}_{t-1} \tag{33}$$

$$m_t = \gamma m^*{}_t + (1-\gamma) m_{t-1} \tag{43}$$

where (43) can be restated as:

$$m^*{}_t = (1/\gamma) m_t - \{(1-\gamma)/\gamma\} m_{t-1} \tag{53}$$

This model implies the estimating equation:[24]

$$\begin{aligned} m_t = {} & a_0\theta\gamma + a_y\theta\gamma y_t + a_R\gamma R_t - a_R\gamma(1-\theta)R_{t-1} + (2-\gamma-\theta)m_{t-1} \\ & -(1-\theta)(1-\gamma)m_{t-2} + \gamma\{\mu_t - (1-\theta)\mu_{t-1}\} \end{aligned} \tag{54}$$

where a_0, $a_y > 0$, $a_R < 0$ and $0 \leq \gamma, \theta \leq 1$. The estimating form of (54) is:

$$m^{\mathrm{d}}{}_t = \alpha_0 + \alpha_1 y_t + \alpha_2 R_t + \alpha_3 R_{t-1} + \alpha_4 m_{t-1} + \alpha_5 m_{t-2} + \eta_t \tag{55}$$

where:

$$\begin{aligned} \alpha_0 &= a_0\theta\gamma \\ \alpha_1 &= a_y\theta\gamma \\ \alpha_2 &= a_R\gamma \\ \alpha_3 &= -a_R\gamma(1-\theta) \\ \alpha_4 &= (2-\gamma-\theta) \\ \alpha_5 &= -(1-\theta)(1-\gamma) \\ \eta_t &= \gamma\{\mu_t - (1-\theta)\mu_{t-1}\} \end{aligned}$$

24 The procedure is: lag each term in (11) by one period and multiply by $(1-\theta)$. Call it (11′). Subtract (11′) from (11) and designate the resulting equation as (11″). Substitute (33) in (11″) to eliminate $y^{\mathrm{p}}{}_t$. This will give $m^*{}_t - (1-\theta)m^*{}_{t-1}$ on the left-hand side and terms in y_t, rather than $y^{\mathrm{p}}{}_t$, on the right-hand side. Now use (43) to substitute for $m^*{}_t - (1-\theta)m^*{}_{t-1}$ and rearrange it as (54).

(55) provides a more general estimating equation than either the PAM model or the adaptive expectations model. These two models are therefore nested in (55), with the PAM one alone obtained when $\theta = 1$ and the adaptive expectations obtained when $\gamma = 1$. Hence (55) provides a way of testing whether there exist both or either of these processes. However, (55) is not necessarily preferable to the alternative estimation procedure that uses the PAM and rational expectations to estimate expected income.

The structural coefficients in (54) are: a_0, a_y, a_R, γ, θ. The coefficients in the estimating equation (55) are: $\alpha_0, \ldots, \alpha_5$. Hence, there are only five structural coefficients compared with six coefficients in the estimated equation, so that appropriate non-linear restrictions have to be imposed on the α_i in the estimating equation (55).

The earlier criticisms of adaptive expectations from the rational expectations perspective also apply here. To reiterate, the criticism is that adaptive expectations are backward looking and ignore information that may be available to the individual about the future as well as on other variables. Further, if the expectations parameter θ shifts, the estimating equation (55) would shift, without it being transparent whether the shift is due to a shift in γ, in θ, or in the coefficients of the demand function. By comparison, using the rational expectations procedures to estimate ${y^{\mathrm{p}}}_t$ in the first step, and then estimating the demand function with PAM, will more clearly disclose the source of the shift in the money demand function.

8.8 Autoregressive distributed lag model: an introduction

Now suppose that the demand for real balances depends on the current and lagged values of real income and its own lagged values. That is, it has the form:

$$m_t = a_0 y_t + a_1 y_{t-1} + a_2 y_{t-2} + \cdots + b_1 m_{t-1} + b_2 m_{t-2} + \cdots \tag{56}$$

This equation is the representation of the general *autoregressive distributed lag (ARDL) model*. y_{t-i} can be replaced by $L^i y_t$, where L^i is the lag operator, which can be treated as a variable subject to mathematical manipulation in the following manner:

$$\begin{aligned} a_0 y_t + a_1 y_{t-1} + a_2 y_{t-2} + \cdots &= a_0 y_t + a_1 L y_t + a_2 L^2 y_t + \cdots \\ &= y_t(a_0 + a_1 L + a_2 L^2 + \cdots) \\ &= a(L) y_t \end{aligned} \tag{57}$$

where $a(L)$ is the polynomial $(a_0 + a_1 L + a_2 L^2 + \cdots)$ in L. Hence, (56) can be rewritten as:

$$m_t = a(L) y_t + b(L) m_t \tag{58}$$

where:

$$a(L) = a_0 + a_1 L + a_2 L^2 + \cdots$$

$$b(L) = b_1 L + b_2 L^2 + \cdots$$

Therefore,

$$m_t - b(L) m_t = a(L) y_t$$

$$m_t = [\{1 - b(L)\}^{-1} \cdot a(L)] y_t \tag{59}$$

so that m_t becomes a function solely of y_t and its lagged terms, without its own lagged values, which are now omitted from the explanatory terms. (59) is the compact form of the ARDL lag model.

An illustration: a simple ARDL model

As an illustration, consider the simplest example of (59) where $a(L) = a_0$ and $b(L) = b_1 L$. That is, (56) is simplified to:

$$m_t = a_y y_t + b_1 m_{t-1} \tag{60}$$

In this case, (59) simplifies to:

$$m_t = \{1 - b_1 L\}^{-1} \cdot a_y y_t \tag{61}$$

Expand $\{1 - b_1 L\}^{-1}$ in a *Taylor's series* around $E(b_1 L) = 0$, where $E(b_1 L)$ is the mean value of $b_1 L$. This gives:

$$\{1 - b_1 L\}^{-1} = \{1 + b_1 L + b_2 L^2 + \cdots$$

Hence, (61) becomes:

$$m_t = \{1 + b_1 L + b_2 L^2 + \cdots\} a_y y_t \tag{62}$$

$$= a_y y_t + a_y b_1 y_{t-1} + a_y b_2 y_{t-2} + \cdots \tag{63}$$

While (60) and (63) are mathematical transformations of each other, so that their economic content is identical, the money demand function in the form of (63) does not contain the lagged value of the endogenous variable, although we started with equation (60) where it does so. Conversely, we could have started with (63) without the lagged money term and derived (60) as equivalent to it. Hence a comparison of (60) and (63) – and of (59) with (56) for the general case – leads to the caution that it may not be possible to distinguish between a money demand equation which contains the lagged values of the endogenous variable and other independent variables, and one which contains only the current and lagged values of the independent variables.

The general ARDL model with the suitable addition of disturbance terms is now in common usage in monetary analysis, and falls in the category of vector autoregression (VAR) models.[25] Its relationship with the now popular cointegration and error-correction estimation is given in the appendix to Chapter 9.

8.9 Demand for money in the open economy

This book has so far concentrated on the demand for money in the closed economy. This is the general pattern of studies on money demand. However, economies are becoming

25 In some of the VAR models used in the analysis of monetary policy, the disturbance terms are interpreted and modeled as policy initiatives, thereby allowing the dynamic intertemporal impact of various policy options to be derived. This use of VAR models has made them fairly popular in the dynamic analysis of monetary policy.

increasingly open to flows of commodities and financial assets, so that a special topic in the money demand literature deals with money demand in the open economy, in which economic units have access not only to domestic financial assets but also to foreign ones.

For portfolio investments in open economies, the financial alternatives to holding domestic money include the currencies and bonds of foreign countries, in addition to domestic bonds, so that the determinants of the domestic money demand should include not only the rates of return on domestic assets but also those on foreign assets. Since these assets include foreign money holdings, money demand studies for open economies need to pay special attention to substitution between domestic and foreign monies. This determination is especially relevant for open economies in which foreign currencies are extensively traded and foreign monies are part of the domestic media of payments. Note that the relevant literature on substitution between domestic and foreign money in the open economy uses the word "currency" for money. This chapter follows this usage.

Currency substitution (CS) can be defined as substitution between domestic and foreign currencies, which is "*currency–currency substitution.*" Substitution can also exist between domestic currency and foreign bonds, and between domestic currency and domestic bonds, which are "*currency–bond substitutions*." Designating, respectively, the nominal values of domestic money, foreign money, domestic bonds and foreign bonds by M, M^*, B and B^*, CS can be measured by $\partial M/\partial M^*$, while the various currency–bond substitutions would be measured by $\partial M/\partial B$, $\partial M/\partial B^*$, $\partial M^*/\partial B$ and $\partial M^*/\partial B^*$, or by their corresponding elasticities.

Giovannini and Turtleboom (1994), Mizen and Pentecost (1996) and Sriram (1999) provide extensive reviews of the CS literature.

8.9.1 Theories of currency substitution

The magnitude of CS will depend both on portfolio selection considerations – since both M and M^* are assets in a portfolio[26] – and on substitution between them as media of payments in the domestic economy. Therefore, the relevant approaches to the degree of CS are the portfolio/asset approach and the transactions approach.

For the asset/portfolio approach, the relevant theory would be the theory of portfolio selection, as set out in Chapter 5, which would treat M and M^* among the assets in the portfolio. This theory would determine substitution between currencies on the basis of their expected yield and risk. Two currencies would therefore be perfect substitutes if they had identical returns. They would be poor substitutes if, with identical risk, the return on one dominated that on the other. This identity of risk dominance does not in general apply in practice. Note that if some types of bonds were riskless, then, with a higher return, bonds would dominate over money, so that there would be zero portfolio demand for currency.

For the transactions approach to the demand for media of payments, it is the general acceptance in daily exchanges and payments that would determine the degree of substitution between the alternative assets. If the foreign currencies do serve as a medium of payments in the domestic economy, the classic demand analysis for the total of the media of payments,

26 If neither domestic nor foreign currency pay interest, substitution between them can only occur because of changes in the expected exchange rate, which in normal circumstances is only a small part of the total return r^* on foreign bonds.

i.e. for the sum of M and M^*, is the Baumol–Tobin inventory analysis presented in Chapter 4. Under this approach, since domestic and foreign bonds do not serve as media of payments they would have a relatively low substitutability with the domestic currency, while that between M and M^* could be much higher. Further, the demand for $(M + M^*/\rho)$[27] would be a function of the domestic expenditures or GDP to be financed. For a given amount of transactions or expenditures to be financed, an increase in one medium of payments implies a decrease in the other, so that transactions demand analysis implies that $\partial M/\partial(M^*/\rho) < 0$. That is, in economies in which both M and M^* do act as media of payments, $\partial M/\partial(M^*/\rho)$ would be negative and significant. In the limit, if domestic residents are indifferent whether they receive payments in the domestic money or in the foreign one, $E_{M,M^*} = -1$, where $E_{M,M^*} = (M^*/M)(\partial M/\partial(M^*/\rho))$. This elasticity would be very much smaller in absolute magnitude, or non-existent, in open economies in which the usage of foreign currency for domestic payments involves significant additional costs to those for payments in the local currency. If this cost is sufficiently high, $E_{M,M^*} = 0$. Therefore, the magnitude of E_{M,M^*} is clearly likely to vary between economies which do not extensively use foreign monies in domestic payments for goods[28] and those economies in which the foreign money is extensively used as a medium of payments, alongside (or in preference to) the domestic money. "Partially dollarized economies" – defined as ones in which the domestic currency and the foreign one circulate side by side, with buyers and sellers indifferent between their use in settling transactions – are especially ones in which E_{M,M^*} tends to -1.[29]

Handa (1988) argued that economic agents in even very open economies but without effective dollarization tend to use the domestic currency as the preferred medium of payments and do not easily switch to the use of foreign currencies for payments because of the transactions costs[30] imposed on retail payments. He therefore designated the domestic currency as being the "preferred habitat"[31] for the domestic medium of payments. Under this hypothesis, the degree of substitutability between the domestic currency and a given foreign one would depend on the latter's acceptance for payments in the domestic economy or the cost and ease of conversion from the latter into the former. In general, there would be a very significant transactions cost in conversion of foreign currencies into the domestic currency. These costs lie in the spread between the purchase and sale conversion rates and in banks' commissions, and are usually quite significant for the size of the transactions of the representative household in the economy. Further, in retail transactions, payment in a foreign currency is usually at an unfavorable exchange rate set by the retailer. Consequently, the general presumption under the preferred habitat approach would be that foreign currencies will have low elasticities of substitution with the domestic currency, except possibly in special

27 M^* is in the foreign currency and needs to be divided by the exchange rate ρ (defined as units of the foreign currency per unit of the domestic currency) to convert it to domestic currency units.

28 This would be so if, when merchants do accept foreign monies, they do so at quite unfavorable exchange rates or impose transactions costs.

29 A fully dollarized economy will have only the US dollar as its circulating medium of payments, without a distinct domestic currency, so that CS cannot be studied for such an economy.

30 These would include rather unfavorable exchange rates used by merchants over the rates of banks, the use by banks of high commissions and large spreads between buying and selling rates, etc.

31 The term "preferred habitat" is borrowed from one of the approaches used to explain the substitution among government bonds of different maturities and has been offered by some economists as an explanation of the term "structure of interest rates." These issues are covered in Chapter 21.

cases where a particular foreign currency is generally accepted in payments at par in the domestic economy. To illustrate, while sellers in Canada often accept US dollars, their offer by buyers is not all that common, because there is a greater cost to paying in the US dollar than is specified by the bank exchange conversion rate. Hence, under the transactions approach, the degree of substitution between the US dollar and the Canadian dollar need not be high and could be quite low.[32] The Canadian dollar finds almost no acceptance in the United States, so they are poor substitutes in the US economy. Further, in the Canadian economy, even if the Canadian and US dollars proved to be good substitutes, British currency is not generally accepted and would be a poor substitute for the Canadian dollar. Most open economies tend to be of this type, so that, except for special cases, the preferred habitat hypothesis implies that we should expect even quite open economies (open but without extensive usage of foreign currencies in domestic retail payments) to have E_{M,M^*} close to zero or with a small negative value.

Among the special cases of possibly high CS was the historical use of the local currency and the imperial one in colonies during the colonial era. Another special case is the use of the US dollar as a second medium of payments in domestic transactions in partially dollarized economies.[33] For such economies, the transactions demand for the media of payments implies that, for a given amount of transactions and GDP to be financed in economies in which both M and M^* act as media of payments, a decrease in one would have to be offset by an equivalent increase in the other. Hence, partially dollarized economies are especially likely to have E_{M,M^*} equal to –1, and an infinite elasticity of substitution,[34] while non-dollarized economies will have significantly lower elasticities of substitution.

Two broad approaches to CS: weak substitutability between monies and bonds

It is an implicit assumption of the CS literature that weak separability (see Chapter 7) exists between the four financial assets (domestic money, foreign money, domestic bonds and foreign bonds) and other goods, which include commodities and leisure, so that the demand functions for these four assets can be estimated by using only the returns on the four financial assets and the amount to be allocated among them. Proceeding further, the literature allows two possibilities:

A. Preferences over the domestic and foreign monies are not weakly separable from domestic and foreign bonds. That is, $U(M^*, M^*/\rho, B, B^*)$ is not weakly separable into a sub-function with M^* and M^*/ρ. Estimations related to this hypothesis have been labeled in the CS literature the "*unrestricted approach.*" As is discussed later, this approach is more suited to the portfolio approach than to the transactions one. In this approach, the demand function for domestic money will include the returns on all four assets, in addition to other variables, such as a scale variable.

32 Handa (1988) used his preferred-currency hypothesis to explain the oft-estimated low degree of CS between the Canadian and the US dollars, even though Canadians are very familiar with the US currency and the United States is Canada's largest trading partner.

33 If there were no banking or other transactions charges between the domestic currency and the foreign one, they would in the limit become perfect substitutes.

34 This will not be so in a fully dollarized economy, which will only have the US dollar as its circulating medium of payments, without a distinct domestic currency, so that CS cannot be studied for such an economy.

B. Preferences over domestic and foreign monies are weakly separable from domestic and foreign bonds. That is, $U(M^*, M^*/\rho, B, B^*)$ is weakly separable into a sub-function with M^* and M^*/ρ, so that:

$$U(M^*, M^*/\rho, B, B^*) = U(f(M, M^*/\rho), B, B^*).$$

Estimations related to this hypothesis have been labeled the "*restricted approach*" in the CS literature. This approach is appropriate for the demand for the two monies as domestic media of payments. It allows the possibility that domestic money and foreign money may act as media of payments in the domestic economy, but bonds do not.[35] If this is so, the demand functions for M and M^* can be estimated as a function of ρ, the returns on M and M^* and the amount to be allocated between them. Such estimation is said to be "restricted," since it is independent of the returns on bonds.

8.9.2 Estimation procedures and problems

There are three common methods of estimating currency substitution. These are:

- estimation of the elasticities of substitution
- estimation of a money demand function
- estimation of the ratio of domestic to foreign money balances.

Estimation of the elasticities of substitution

This procedure involves estimation of the Euler equations (first-order conditions) based on a constant (CES) or variable (VES) elasticity-of-substitution function. This method follows Chetty's procedure, explained in Chapter 7. In the unrestricted choice framework, the domestic money and foreign money balances, along with domestic and foreign bonds, would appear in the VES utility function. The estimating equations will be derived from the Euler conditions (see Chapter 7).[36] This procedure allows estimation of the elasticity of substitution between the two monies, and between the domestic money and the two types of bonds. The estimating equations in this case would be similar to those specified in Chapter 7 for the VES model, and are not listed here explicitly. In the restricted choice framework, with domestic and foreign monies weakly separable from bonds, the VES function would only include the two monies, so that the foreign money holdings of domestic residents would be regressed on domestic money balances and their "price" ratio. Among the studies based on this approach are those of Miles (1978) and Handa (1971).

35 That is, in effect, investors treat domestic and foreign monies as weakly separable (see Chapter 7) from domestic and foreign bonds. Further, if the two monies are weakly separable from bonds and the two types of bonds are weakly separable from the two monies, the preference/utility function over them would have the form:

$$U(M^*, M^*/\rho, B, B^*) = U(f(M, M^*/\rho), g(B, B^*))$$

36 These would properly include both currency and deposits, held by domestic residents at home or abroad. However, the data on foreign currency holdings is rarely available. Further, the data on the foreign demand deposits of domestic residents, held in foreign countries, is often not available. These omissions can very significantly affect the empirical findings.

Note that an alternative to the VES model is to assume a priori a unit elasticity of substitution between the assets and construct their chained Divisia or certainty-equivalence index (for which, see Chapter 7) with time-variant expenditure shares. These methods can be used for M and M^* only under the weak separability assumption of the restricted choice framework, or for all four financial assets for unrestricted choice. Estimation is not needed for the construction of the Divisia and certainty-equivalence aggregates.

Estimation of the domestic money-demand function

This estimation procedure is to expand the estimating money-demand equation to include among its regressors the return on at least one foreign currency, as well as returns on foreign bonds (and sometimes also physical) assets. This is the more common method of estimating currency substitution. It can be found in Bordo and Choudhri (1982), Bana and Handa (1987)[37] and Handa (1988).

For the unrestricted choice approach, the standard money-demand function, modified to take account of foreign currencies and foreign bonds as alternatives to domestic money, is usually specified as:

$$m^{\mathrm{d}} = \alpha_0 + \alpha_R R + \alpha_y y + \alpha_\varepsilon \varepsilon^{\mathrm{e}} + \alpha_{R^*} R^* + \mu \qquad (64)^{38}$$

where:

m^{d} = domestic money balances in real terms
y = domestic real national income
R = nominal yield on domestic bonds (= domestic rate of interest)
R^{F} = nominal interest rate on foreign bonds
R^* = nominal yield on foreign bonds
(= foreign rate of interest + expected appreciation of the foreign currency)
ρ = exchange rate (domestic currency per unit of foreign currency)
ε^{e} = expected return on the foreign currency
μ = disturbance term.

In (64), the three rates of return are ε, R and R^*. Note that the returns on domestic and foreign currencies include both their non-monetary returns – that is, their liquidity services, etc. – and the change in their nominal values relative to each other. While the liquidity and other non-monetary services are often critical for the demand for foreign currencies, data on them is usually non-existent, so that they are almost always excluded from the analysis. This is a significant deficiency of the empirical studies on currency substitution since, except in effectively dollarized economies, the acceptance in exchanges of domestic and foreign currencies and the ease of payment differ considerably.

37 This study argued that the degree of currency substitution would differ between fixed and flexible exchange rates, and found higher substitution under flexible rates.

38 The nominal exchange rate ρ is sometimes added as another explanatory variable in equation (64) since domestic money is in the domestic currency unit while the foreign monies and bonds are in foreign currency units, so that the latter need to be converted into the domestic currency. Some studies replace ρ by ρ^{r} (the real exchange rate), which is the rate of exchange between the domestic currency and foreign commodities, if the emphasis of the model is on the medium-of-payments role of monies in purchases of commodities. Under our definition of ρ, $\rho^{\mathrm{r}} = \rho P/P^*$. However, empirically, PPP does not usually hold, so it should not be assumed.

There would be similar demand functions for foreign money, domestic bonds and foreign bonds.

Forced by the lack of data on the non-monetary/liquidity costs of domestic and foreign monies, the monetary return on foreign currencies is measured by the expected rate of appreciation of the foreign currency *vis-à-vis* the domestic currency.[39] This expected appreciation equals $(-\partial\rho/\partial t)^{e}$, where ρ is the number of units of the domestic currency per unit of the foreign one, so that $(-\partial\rho/\partial t)^{e}$ is the opportunity cost of holding the domestic currency rather than the foreign one. Therefore, ε^{e} in (64) is measured by $(-\partial\rho/\partial t)^{e}$. In empirical estimations using quarterly data, the proxy usually used for ε^{e} is $(F-S)/S$, where F is the 90-day forward exchange rate and S is the spot rate. For empirical studies on countries other than the USA, the foreign currency is usually taken to be the US dollar.

In open economies with perfect financial markets, the domestic and foreign interest rates are related by the interest rate parity (IRP) condition:

$$(1+R_t)=(1+R^{F}{}_t)(1+\varepsilon^{e}{}_t) \tag{65}$$

where R^{F} is the rate of interest on foreign bonds and $\varepsilon^{e}(=(\partial\rho/\partial t)^{e})$ is the expected rate of depreciation of the domestic currency. The common approximation to (65) is:

$$R_t=R^{F}{}_t+\varepsilon^{e}{}_t \tag{66}$$

R_t, R^{f}_t and $\varepsilon^{e}{}_t$ are all arguments of the open-economy money demand function. (66) implies that only two of these three variables are independent of each other, so that any two of them, but not all three, should be included in the estimating money demand equation. The two variables so selected are usually the domestic rate of interest and the expected exchange rate appreciation: many studies set $\alpha_{R^*}=0$ prior to estimation, using the intuition that substitution between domestic money and foreign bonds is likely to be minimal.

(64) is usually estimated in a log-linear form, so that its coefficients represent elasticities. The cross-price elasticity α_R is the indicator of price-substitution[40] between the domestic currency and domestic bonds, and the cross-price elasticity α_{R^*} is the indicator of price-substitution between the domestic currency and foreign bonds.[41] α_ε is the cross-price elasticity[42] with respect to the return on foreign currencies and can therefore be used as an indicator of CS.

In a four-asset portfolio selection analysis, the signs of any of the three cross-elasticities α_ε, α_R and α_{ρ^*} are not specified by theory and could be negative or positive.[43] An empirically determined negative sign is interpreted as evidence of substitution between the domestic currency and the relevant asset. While a positive sign is sometimes interpreted as evidence of complementarity, this interpretation is not necessarily correct since it can reflect some

39 The expected return is the expected rate of depreciation less the cost of holding foreign currencies. This cost can be very significant where it is illegal to hold and/or deal in foreign currencies, but is minimal in an unregulated free financial system.

40 Note that the elasticity of domestic money demand with respect to $(\partial\rho/\partial t)^{e}$ is a "price" elasticity and as such does not directly measure the elasticity of substitution between the domestic and the foreign currency. The relationship between the price elasticity and the elasticity of substitution has already been discussed in Chapter 7 in connection with Chetty's procedure.

41 Cuddington (1983) argued that a significant coefficient of the return on foreign bonds is evidence of capital movement, not CS, while CS requires the coefficient of ε to possess a negative sign and be significant.

42 The cross-return elasticity is a cross-elasticity with respect to a change in the price (in our context, return) of another variable.

43 This sign would be specified only as negative in a two-asset model.

other effect. This possibility is especially likely for the sign of α_{R*}, as we illustrate later through the discussion on the substitution between M and M^* in the medium-of-payments role.

Estimation of capital mobility

Capital mobility is distinct from CS and may be defined as the net outflow of funds from the domestic economy into foreign assets, so that it would be specified by the overall substitution between the sum of the domestic currency and domestic bonds into foreign currency and foreign bonds. This would require the estimation of both the domestic money and bond equations. Therefore, the coefficients in the money demand function alone cannot be used as an indicator of capital mobility.[44]

8.9.3 The special relation between M and M in the medium-of-payments function*

Since the domestic currency and foreign bonds are both assets, portfolio selection theory implies that an increase in the return on foreign bonds relative to the return on the domestic currency (the riskless asset with a zero return) would cause substitution between them, thereby implying that $\partial M/\partial R^* = \alpha_{R^*} \leq 0$. This effect can be decomposed into two components specified by:

$$\frac{\partial M^{\mathrm{d}}}{\partial R^*} = \left[\frac{\partial (M^{\mathrm{d}})}{\partial R^*}\right]_{M^*=M^*} + \frac{\partial M^{\mathrm{d}}}{\partial (M^*/\rho)} \frac{\partial (M^*/\rho)}{\partial R^*} \tag{67}$$

In (67), the first term on the right represents direct substitution between M and foreign bonds, holding foreign money balances constant. This (direct) effect occurs because the increase in foreign bonds increases the opportunity cost of holding domestic money relative to foreign bonds. The second term on the right is an indirect effect occurring through $\partial M^*/\partial R^*$, which arises because an increase in R^* also increases the opportunity cost of holding foreign money. As these balances decrease, the public has to increase domestic money balances in order to arrive at the desired level of the overall media of payments needed to finance its expenditures.

The direct effect is primarily determined by portfolio selection, which treats domestic money as an asset held for its return relative to other assets. Except in conditions where the domestic bonds do not exist or their security is doubtful, the significant portfolio switch caused by a rise in R^* is likely to be between foreign bonds and domestic bonds (rather than domestic money) since both are income-earning assets. It is not likely to be between foreign bonds and domestic money. Therefore, while portfolio selection analysis implies that $(\partial M^{\mathrm{d}}/\partial R^*)|_{\Delta M^*=0} \leq 0$, this direct effect is likely to be quite weak in normal financial conditions.

The indirect effect in (67) is the multiple of two elements, $\partial M^{\mathrm{d}}/\partial (M^*/\rho)$ and $(\partial M^*/\rho)/\partial R^*$. On the second of these two elements, both the portfolio selection and the transactions demand analyses imply that $(\partial M^*/\rho)/\partial R^* < 0$. On the first element, $\partial M^{\mathrm{d}}/\partial (M^*/\rho)$ is the substitution

44 This is so even if ε is omitted from the regression. In this case, the coefficient of $R^*(= R^{\mathrm{F}} + \varepsilon)$ in the estimation of M^{d} is likely to capture both CS and the substitution between domestic currency and foreign bonds.

between domestic and foreign monies, which constitutes CS. Our earlier discussion on this point implies that $\partial M^{d}/(M^{*}/\rho) \leq 0$. Hence, in (67), the second term on the right is non-negative. Therefore:

$$\left[\frac{\partial M^{d}}{\partial R^{*}}\right]_{M^{*}=\overline{M^{*}}} \leq 0, \tag{68}$$

$$\frac{\partial M^{d}}{\partial (M^{*}/\rho)} \leq 0 \quad \text{and} \quad \frac{\partial M^{*}/\rho}{\partial R^{*}} \leq 0 \tag{68'}$$

so that

$$\frac{\partial M^{d}}{\partial (M^{*}/\rho)} \frac{\partial (M^{*}/\rho)}{\partial R^{*}} \geq 0 \tag{68''}$$

Since, from (68), the first term (the direct effect) on the right-hand side of (67) is non-positive and the second one (the indirect effect), from (68″) is non-negative, the sign of $\partial M^{d}/\partial R^{*}$ in (67) – and hence of $\alpha_{R^{*}}$ in (64) – is analytically indeterminate and will depend on the relative magnitudes of the direct and indirect effects. Assuming that the major substitution in portfolio selection is unlikely to occur between the domestic money and foreign bonds,[45] $(\partial M^{d}/\partial R^{*})|_{\Delta M^{*}=0}$ in (67) is likely to be relatively small, so that our hypothesis is that the second term on the right-hand side of (67) will dominate the sign and magnitude of $\partial M^{d}/\partial R^{*}$ for most economies.

Now focusing on the magnitude of $\partial M^{d}/\partial (M^{*}/\rho)$, our earlier discussion of this CS effect implies that it will be relatively weak in economies in which the foreign money is not extensively used as a medium of payments in domestic transactions.[46] However, in economies in which foreign monies function as one of the domestic media of payments, $\partial M^{d}/\partial (M^{*}/\rho)$ will be negative and significant. Therefore, in the context of equation (64), if both the domestic and foreign currencies are widely used as media of payments in the domestic economy and the demand for the overall media of payments is determined by domestic GDP, an increase in the return on foreign bonds could decrease the holdings of the foreign currency in the domestic economy, which need to be balanced by an increase in the domestic money balances. That is, the increase in R^{*} would induce a significant increase in M^{d}, so that the second term in (67) is likely to dominate and $\alpha_{R^{*}}$ in (64) should be positive.

Hence, our hypothesis for partially dollarized economies is that $\alpha_{R^{*}}$ should be positive and significant. By comparison, for economies in which foreign money is not in extensive usage as one of the media of payments, the first term on the right-hand side of (67) is likely to dominate, so that $\alpha_{R^{*}}$ in (64) should be insignificant or negative.

45 Conversely, the probable substitution is likely to be between domestic and foreign monies, between domestic money and domestic bonds, and between foreign money and foreign bonds.

46 For such economies, Handa (1988) specified the domestic currency as the preferred habitat for the domestic medium of payments, from which substitution into the foreign currency would be relatively low.

Estimation of $M/(M^*/\rho)$

The third procedure for estimation of the money-demand function focuses on the estimation of M/M^* or of $M/(M^*/\rho)$. We have argued above that there is a special relationship between M and M^* because of their substitution in domestic payments. This could be captured by a weakly separable preference function over M and M^* (as in Bordo and Choudhri, 1982) or a "monetary services production function" (as in Ratti and Jeong, 1994). For such functions, CS can also be assessed by estimating the function for the ratio $M/(M^*/\rho)$ rather than the demand function for domestic money only. In the general, unrestricted, case, this ratio will also be a function of the explanatory variables in (64), so that the corresponding log-linear equation with all variables in logs would be:

$$M^{\mathrm{d}}/(M^{*\mathrm{d}}/\rho) = f(\varepsilon^{\mathrm{e}}, R, R^*, Y, \rho) = \beta_0 + \beta_\varepsilon \varepsilon^{\mathrm{e}} + \beta_R R + \beta_{R^*} R^* + \beta_Y Y + \beta_\rho \rho \tag{69}$$

As pointed out earlier, the approximate form of the IRP hypothesis implies that ε^{e}, R and R^* are linearly related for an economy with a high degree of capital mobility, so that ε^{e} can be left out of the explanatory variables. β_R^* is likely to be positive for economies in which foreign money is significant as a medium of payments in transactions. However, our conjecture is that $\beta_{R^*} \leq 0$ for other economies in which the foreign money does not circulate extensively in domestic payments because, among other reasons, retailers do not give the bank rate of exchange and/or charge commissions.

8.9.4 Other studies on CS

In their estimating equation for CS, Ratti and Jeong (1994) claim to combine a "dynamic monetary services" model with portfolio allocation. Their model[47] implies that the optimal ratio $M/(M^*/\rho)$, under purchasing power parity and interest rate parity, equals $(\rho \cdot P/P^*)(R/R^{\mathrm{F}}(1+\varepsilon))$, so that, in log-linear term:

$$M/(M^*/\rho) = \beta_1(\rho \cdot P/P^*) + \beta_2(R/R^{\mathrm{F}}(1+\varepsilon)) \tag{70}$$

where P^* is the foreign price level, $(\rho \cdot P/P^*)$ is the real exchange rate, which is included since foreign money balances need to be converted to their purchasing power over domestic commodities, and $(R/R^{\mathrm{F}}(1+\varepsilon))$, with $R^* \equiv R^{\mathrm{F}}(1+\varepsilon)$, is the relative rate of return on domestic and foreign bonds. Note that if absolute purchasing power parity (PPP) holds, $(\rho \cdot P/P^*) = 1$, so that, in (70), $M/(M^*/\rho) = (R/R^*(1+\varepsilon))$, which implies that, in a regression, β_1 should not be significantly different from zero. If it is, the PPP hypothesis is rejected. But if IRP holds, then $(R/R^*(1+\varepsilon))$ equals unity, so that $M/(M^*/\rho) = (\rho \cdot P/P^*)$, implying that β_2 should not be significantly different from zero; if it is, the PPP hypothesis is rejected. If both PPP and IRP hold, then $M/(M^*/\rho) = 1$, implying that both β_1 and β_2 should not be significantly different from zero. Further, if neither PPP nor IRP holds, (70) implies that the coefficients of both $(\rho \cdot P/P^*)$ and $(R/R^*(1+\varepsilon))$ should be unity. Therefore, (70) seems to represent a very restrictive model, whose value lies not so much in providing estimates of CS but rather whether or not either or both PPP and IRP hold.

47 For the details of this model, we refer the reader to Ratti and Jeong (1994). This model does not include money directly in the utility function but relies on a "transactions function" that specifies the real resources needed for transactions, with "money services" being provided by domestic and foreign money holdings.

In economies in which domestic residents have limited or no access to foreign bonds, the return on foreign bonds will not enter the domestic money demand function, so that multicollinearity between changes in the expected exchange rate and returns on domestic and foreign bonds will not pose a problem. This makes the estimation of CS, and the evaluation of its extent, by the estimated coefficient of the expected exchange rate change more credible. For such a context, De Freitas and Veiga (2006) study CS in the context of six Latin American economies.[48] They use a stochastic dynamic optimizing model in which money reduces the losses due to frictions in commodity exchanges. They report evidence of CS for Colombia, the Dominican Republic and Venezuela but not for Brazil and Chile, with ambiguous results for Paraguay.

Conclusions

This chapter has examined the form of the money demand function to be used for estimation. One of these uses expected income as its scale variable, while another uses permanent income. Neither is observable, so that a procedure has to be adopted for their estimation. The rational expectations hypothesis (REH) was proposed for the estimation of expected income, while the adaptive expectations hypothesis was proposed for the estimation of permanent income. Of these, adaptive expectations are backward looking and ignore information that may already be available on the future, but do provide a better measure of permanent income, which is the average expected level of income for the future rather than expected income for the next period.

This chapter also looked at the use of partial adjustment models (PAMs). These models are based on the notion that there are various costs of adjusting money balances quickly, and imply the specific order of the partial adjustment model. The general autoregressive distributed lag model nests the PAM and the adaptive expectations models. An alternative to such a model would be a PAM model with a separate procedure for the rational expectations estimation of expected income.

The open-economy form of the money demand equation distinguishes between currency substitution (i.e. substitution between the domestic and foreign currencies) and capital mobility, which is mainly substitution between domestic and foreign bonds. There are basically three procedures for estimation of currency substitution. These are the estimation of a money demand function, a variable elasticity of substitution function and estimation of the ratio of domestic to foreign currency holdings.

While portfolio theory seems to imply that there should be considerable and increasing currency substitution among the highly open modern economies, the econometric evidence remains quite mixed. This could be due to the preferred habitat role of domestic money balances as the domestic medium of payments. For most economies, foreign monies do not commonly circulate in the economy because of "brokerage costs" imposed by retailers on payments in foreign currencies. However, these costs tend to be trivial for a specific foreign money, which is often the US dollar, in partially dollarized economies, so that such a money functions as a domestic medium of payments in addition to the domestic money. In this case, there should be a high degree of substitution between domestic money and

48 These authors start with an intertemporal relative risk aversion utility function over consumption in different periods and assume that the time spent shopping per unit of consumption expenditures depends on the domestic and foreign monies held.

foreign money.[49] Note that a fully dollarized economy will not have a distinct domestic money.

Summary of critical conclusions

- The appropriate scale variable for the demand for money may be current income, expected income or permanent income, the last one being a substitute for wealth.
- The rational expectations hypothesis is more suitable than adaptive expectations for estimating expected income for the period ahead. For this hypothesis, the unanticipated component of income is usually estimated as income less the statistical estimate of expected income.
- The adaptive expectations approach is the more appropriate statistical method for estimating the average expected income over the future, i.e. permanent income.
- Partial adjustment models provide a way of capturing the lagged adjustment of actual to desired money demand.
- In the open economy, currency substitution (CS) also affects the demand for domestic money. CS is distinct from capital mobility.

Review and discussion questions

1. Are there significant costs in adjusting actual to desired money balances or in changing balances between periods? Or are any such costs relatively insignificant but the delay in the adjustment of actual to desired balances occurs as a consequence of the costs of adjusting commodities and other financial assets to their desired levels? Discuss.
2. Discuss the justification for the use of permanent income in a money demand function. Assuming that adaptive expectations are to be used to derive permanent income, derive the estimating money demand equation.

   Discuss the suitability of rational expectations for estimating permanent income.
3. Discuss the justification for the use of expected income in a money demand function. Assuming that rational expectations are to be used to derive expected income, derive the estimating money demand equation.

   Discuss the suitability of adaptive expectations for estimating expected income.
4. Starting with a cost function leading to a second order partial adjustment model, and a money demand function with expected income as the scale variable, derive the appropriate form of the estimating equation for the demand for money.
5. Starting with a cost function leading to a second order partial adjustment model, and a money demand function with permanent income as the scale variable, derive the appropriate form of the estimating equation for the demand for money.
6. Define currency substitution and distinguish it from capital mobility, as well as from substitution between domestic and foreign bonds. How are the returns on foreign monies and foreign bonds determined? What multicollinearity problems arise in the open economy money demand equation and how would you choose to resolve them?

49 Ko and Handa (2006) estimate CS in Hong Kong, a partially dollarized economy in which the HK dollar and several foreign currencies (US dollar, British pound and the Chinese yuan) circulate in domestic payments, and report a very significant degree of CS.

7. Define weak separability. Discuss the role that it plays in specifications of the estimation equations for CS?
8. Present and discuss the three major modes of estimating currency substitution for an economy.

References

Arrow, K. "The future and the present in economic life." *Economic Inquiry*, 16, 1978, pp. 157–69.

Bana, I.M., and Handa, J. "Currency substitution: a multi-currency study for Canada." *International Economic Journal*, 1, 1987, pp. 71–86.

Bordo, M.D., and Choudhri, E.U. "Currency substitution and the demand for money: some evidence for Canada." *Journal of Money, Credit and Banking*, 14, 1982, pp. 48–57.

Cuddington, J. "Currency substitution, capital mobility and money demand." *Journal of International Money and Finance*, 2, 1983, 111–33.

Cuthbertson, K. *The Supply and Demand for Money*. London: Blackwell, 1985, Ch. 3.

De Freitas, M.L., and Veiga, F.J. "Currency substitution, portfolio diversification, and money demand." *Canadian Journal of Economics*, 39, 2006, pp. 719–43.

Friedman. M. "The quantity theory of money – a restatement." In M. Friedman, ed., *Studies in the Quantity Theory of Money*. Chicago: Chicago University Press, 1956, pp. 3–21.

Giovannani, A., and Turtleboom, B. "Currency substitution." In Frederick van der Ploeg, ed., *The Handbook of International Macroeconomics*. Oxford: Blackwell, 1994.

Handa, J. "An empirical study of financial intermediation in Canada." *Journal of Financial and Quantitative Analysis*, 4, 1971, pp. 583–600.

Handa, J. "Substitution among currencies: a preferred habitat hypothesis." *International Economic Journal*, 2, 1988, pp 41–62.

Kantor, B. "Rational expectations and economic thought." *Journal of Economic Literature*, 17, 1979, pp. 1422–41.

Ko, K.W., and Handa, J. "Currency substitution in a currency board context: the evidence for Hong Kong." *Journal of Chinese Economic and Business Studies*, 4, 2006, pp. 39–56.

Miles, M.A. "Currency substitution, flexible exchange rates and monetary independence." *American Economic Review*, 68, 1978, pp. 428–36.

Mizen, P., and Pentecost, E.J. "Currency substitution in theory and practise." In P. Mizen, and E.J. Pentecost, eds, *The Macroeconomics of International Currencies*. Cheltenham, UK: Edward Elgar, 1996, pp. 8–43.

Muth, J.F. "Rational expectations and the theory of price movements." *Econometrica*, 29, 1961, pp. 315–35.

Ratti, R.A., and Jeong, B.W. "Variation in the real exchange rate as a source of currency substitution." *Journal of International Money and Finance*, 13, 1994, pp. 537–50.

Sriram, S.S. "Survey of literature on demand for money: theoretical and empirical work with special reference to error-correction models." *International Monetary Fund Working Paper* No. 64, 1999.

9 The demand function for money

Estimation problems, techniques and findings

This chapter presents the estimating function for money demand, an introduction to the appropriate econometric techniques and a summary of the empirical findings on money demand. On the econometric techniques, a major part of the presentation is on cointegration techniques with error-correction modeling for estimating the short-run and the long-run demand for money.

The empirical evidence clearly confirms the dependence of the demand for money on both a scale variable and an interest rate. The issue of which scale variable should be used – current income, permanent income or wealth – is still not settled.

Key concepts introduced in this chapter

- Multicollinearity
- Serial correlation
- Stationarity
- Order of integration
- Unit roots
- Cointegration
- Error-correction modeling

The preceding chapters specify the theoretical analyses of money demand and the general nature of the money demand function. This chapter examines the econometric problems and techniques associated with its estimation, and presents the findings of some of the relevant empirical studies.

A very large number of empirical studies on the money demand function have been published. It would take up too much space to review even the more important of these studies, or to do justice to the ones from which we adopt the results. Among the many excellent reviews of these studies in the literature are those by Cuthbertson (1991), Goldfeld (1973), Feige and Pearce (1977), Judd and Scadding (1982), Goldfeld and Sichel (1990), Miyao (1996) and Sriram (1999, 2000). We shall present only the generic findings on the more significant issues, especially those on the income and interest rate elasticities and the appropriate measure of the monetary aggregate. We intend to pay particular attention to the findings from studies using cointegration and error-correction analysis.

The empirical findings on monetary aggregation reported in Chapter 7 complement the material in this chapter. In particular, the empirical findings concerning the Divisia and certainty equivalence aggregates versus simple-sum aggregates are to be found in Chapter 7 rather than this chapter.

Section 9.1 presents a historical review of money demand estimation and its findings. Sections 9.2 to 9.7 discuss some of the econometric problems that can arise with the data and present the cointegration and error-correction techniques. Section 9.8 presents the findings of some empirical studies using these procedures. Section 9.9 touches on causality. Section 9.10 provides an illustration of the shifts in income and interest-rate elasticities due to financial innovations. Section 9.11 focuses on the desperate search for a stable money demand function.

9.1 Historical review of the estimation of money demand

By the end of the 1960s, the basic form of the money demand function had evolved as:

$$m^{\mathrm{d}} = a_0 + a_R R + a_x x \tag{1}$$

where x is a scale variable. The stochastic form of this function was estimated in either a linear or a log-linear form. During the 1960s, the main disputes were whether money should be defined as M1, M2 or by a still wider definition, whether the interest rate should be short-term or long-term, and whether the scale variable x should be income, permanent income or wealth. The data usually used for estimation was annual.

The 1970s were a period of increasing deregulation of the financial system, with financial institutions offering a variety of interest-bearing checking accounts and checkable savings accounts. There was increasing use of quarterly data in this decade and of the partial adjustment model discussed in Chapter 8. The latter justified the use of the lagged value of money among the explanatory variables, so that the linear or log-linear form of the commonly estimated money-demand function was:

$$m_t^{\mathrm{d}} = a_0 + a_r r_t + a_y y_t + (1-\gamma) m_{t-1} + \mu_t \tag{2}$$

where γ was the adjustment parameter and μ was a white-noise disturbance term.

In an attempt to eliminate serial correlation, a common problem, in money-demand functions or to incorporate a partial adjustment model, (2) was often estimated in its first difference form. The empirical estimates still indicated the stability of the money demand function, but M1 now often, though not always, performed better than M2 and broader aggregates. The value of the adjustment parameter γ in (2) tended to be roughly between 0.20 and 0.5, so that full adjustment to long-run values occurred in about two to six quarters. There was a low impact (one-quarter) real income elasticity (about 0.2) and long-run income elasticity less than 1 (often around 0.7), and a low impact interest elasticity (about -0.02 or smaller) and a long-run interest elasticity roughly between -0.05 and -0.15.[1] The empirical findings on the income and interest elasticities of money demand in Canada were roughly similar.

1 For a summary up to about 1990 of the empirical findings on money demand, see Goldfeld and Sichel (1990).

Income and wealth in the money demand function

The period of the 1950s and early 1960s in many countries was one during which the regulatory authority did not allow interest to be paid on demand deposits. Further, the interest rates paid on savings deposits were subject to upper limits, savings deposits could not be drawn upon by check and a switch from savings deposits to demand deposits often required a personal visit to the relevant financial institution. Under these conditions, the general finding among the empirical studies was that M2 did better than either M1 or measures broader than M2. The explanatory variables that usually performed best with M2 as the dependent variable were medium- or long-term interest rates, with wealth or permanent income as the scale variable. The estimating function was normally stable.

For the data covering the 1950s and 1960s in the USA, regression analysis of the demand for money from equations containing both income and wealth, as well as from equations containing only one of these variables, showed that wealth provided a more stable demand function for money than current income and that, when both variables were included simultaneously, the coefficient of the income variable was insignificant. Permanent income similarly performed better than current income. These results held especially if money was defined as M2 or M3 but not as often in studies where the dependent variable was M1. Further, among functions using income, non-human wealth and permanent income, the empirical estimates showed that functions using a wealth concept gave more accurate predictions of the velocity of circulation of money broadly defined – but not as often for M1– than did those containing current income.

The findings on the economies of scale were uneven. Studies using M1 as the dependent variable often found income elasticities to be less than one, typically around 0.7 or 0.8. Higher income elasticities were usually reported for M2, with some in excess of unity. The reason for this divergence hinges on the inclusion of interest-earning savings deposits in M2. The demand for savings deposits is likely to reflect more strongly a portfolio demand than does M1, so that, with income and wealth positively correlated, the income elasticity of M2 will tend to capture to a greater extent the impact of wealth on savings deposits than does the income elasticity of M1. This portfolio demand could make them a "superior good" for households who experience wealth increases during the sample period.

As between the partial adjustment model and adaptive expectations, with US annual data for 1915–63, Feige (1967) used permanent income as the scale variable and reported instantaneous adjustment. However, Goldfeld (1973), with quarterly US data, found less than instantaneous adjustment. In general, during the 1970s, studies using quarterly data provided evidence of both adaptive expectations and partial adjustment.

Interest rates in the money-demand function

There are many interest rates in the economy, ranging from the return on savings deposits in banks and near-banks to those on short- and long-term bonds. Near-money assets such as savings deposits in commercial banks proved to be the closest substitutes for M1, so that their rate of return seems to be the most appropriate variable as the interest cost of using M1.

But if a broader definition of money were used, the interest rate on medium-term or long-term bonds would become more appropriate (the alternative to holding M2 or M3 is longer term bonds), since the savings components of the broad definition of money themselves earn an interest rate close to the short rate of interest.

The interest rates usually used in estimating money demand are: the interest rate paid on savings deposits in commercial banks, or on those in credit unions (such as Mutual Savings Banks and Savings and Loan Associations in the USA, Caisses Populaires in Quebec, Canada); the yield on Treasury bills or on short-term prime commercial paper and the yield on longer term bonds, such as 3 to 20-year government or commercial bonds. Each of these interest rates seems to perform fairly well, sometimes better and sometimes worse than others, in some study or other, and yields different coefficients.

A uniformly good performance, irrespective of which of the interest rates is included in the regression, is an indication that the various interest rates are related, moving up or down in a consistent pattern, so that it is immaterial which interest rate is included. One theory that points towards such consistency of pattern is the expectations hypothesis on the term structure of interest rates, i.e. on the yields on assets differing in maturity. Chapter 20 presents this hypothesis for the financially well-developed financial markets, pointing out that it has done remarkably well in explaining the differences in the yields of assets differing in maturity. A consequence of such a relationship among interest rates is that the inclusion of more than one interest rate results in multicollinearity and therefore in biased estimates of their coefficients.

However, while the relevant interest rates are closely related, they do not move so closely together that any of them will do equally well in estimation, so that usually one or two of them have to be chosen on empirical grounds for inclusion as regressors. On the wider question of whether the demand for money depends on interest rates or not, there is substantial evidence that the demand for money does depend negatively upon the rate of interest in financially developed economies. This is also the finding of many studies on the less developed countries (LDCs).

Some studies on the LDCs, however, do not find significant interest rate elasticities for a variety of reasons, including regulatory limits on the interest rates in the economy and inadequate access to banking and other financial facilities. In these cases, very often the rate of inflation rather than the published data on interest rates yields better empirical results. This occurs because the regulated interest rates usually do not accurately reflect the expected rate of inflation, as market-determined rates do in developed financial markets, so that land, inventories and other real assets, whose prices better reflect the rate of inflation, become more attractive alternatives than bonds for holding cash.

Various empirical studies have reported that the interest elasticity of money demand is definitely negative and significant, and in the range -0.15 to -0.5.

Money demand and the expected rate of inflation

One of the alternatives to holding money is commodities, which have the (expected) rate of return equal to the expected rate of inflation less their storage and depreciation costs. Some of the commodities – as for example, untaxed plots of land – have minimal storage and depreciation costs, so that the (expected) rate of return on commodities is usually taken to be proxied by the expected rate of inflation. Therefore, the expected rate of inflation is one of the arguments in the money-demand function, in addition to interest rates, as Friedman's analysis of money demand in Chapter 2 pointed out.

However, in perfect financial markets, for small values of the real interest rate and expected inflation, the nominal and the real rates of interest are related by the Fisher equation:

$$R_t = r_t + \pi_t^e \tag{3}$$

where R is the nominal rate of interest, r is the real one and π^e is the expected inflation rate. At significant rates of inflation, variations in the real rate tend to be much smaller in magnitude than the expected inflation rate, so that R_t and π^e_t will be closely correlated. Given this close correlation and that between π^e_t and the actual rate of inflation π_t, R and π also tend to be closely correlated in periods with significant inflation rates. Therefore, incorporating both R_t and π_t in the money demand equations often leads to multicollinearity and biased estimates of their coefficients. As a way around these statistical problems, π_t is often dropped in favor of R_t from the estimated money demand equations for developed economies with market determination of R_t. However, economic theory implies its inclusion in addition to the inclusion of interest rates, so that its omission could result in a misspecified equation.

In economies such as the LDCs', where the financial markets are not well developed, ceilings are often imposed on the rates of interest that can be legally paid and there could exist both an official interest rate and a free or black market rate. Further, reliable data on interest rates may not be available. In these cases, π^e_t should be retained in the estimating equation in addition to – and sometimes even to the exclusion of – the interest rate. Note that the proper variable is π^e_t and that π_t is only one of the possible proxies to it.[2]

The liquidity trap

One of the questions of interest in monetary theory since the time of Keynes, discussed in Chapter 2 above, has been about the empirical existence of the liquidity trap. Keynes posited the possible existence of such a trap, though he also expressed the belief that he did not know of any case where it had existed.

One possible method of testing for the existence of the liquidity trap is to estimate the demand for money separately for periods with differing ranges of the prevailing interest rates. Estimates showing that the interest elasticity of demand tends to increase in periods with lower ranges of interest rates, and especially those showing a substantial increase at very low interest rates, can be interpreted as raising a presumption that the liquidity trap could have existed empirically. However, empirical studies so far have not revealed such a pattern. Velocity functions estimated separately for each decade did not find any higher interest elasticity of the demand for money during the 1930s, when interest rates were low than during other decades with higher interest rates. Further, regressions incorporating data from the 1930s did fairly well in predicting velocity during the subsequent decades, implying that the interest elasticities during the 1930s did not differ substantially from those of more normal conditions. These studies indicate that the liquidity trap does not seem to have existed in the US economy for any significant period, if at all, during the Great Depression of the 1930s and is even less likely a possibility for other periods.

Theoretically, the liquidity trap should come into existence if the nominal yield on bonds becomes zero. The Japanese economy in recent decades provides an interesting experiment on the liquidity trap since it has had short-term interest rates close to zero. Bae *et al*. (2006) have studied different money demand functions for Japan using linear and non-linear cointegration techniques with quarterly data from 1976:1 to 2003:4. They report that the interest elasticity for their various monetary aggregates, including M1, is much higher at low interest rates than

2 If the real rate of interest were constant, the nominal rate of interest would become the better proxy for π^e_t.

at higher rates, thereby favoring the conjecture that the liquidity trap may exist at interest rates that are zero or close to zero.

Shifts in the money demand function

Much greater impact of financial deregulation was felt in the 1980s than had been permitted or achieved in the 1970s. Further, technological and product innovation in the financial sector was very rapid. Computers also came into general use in firms and households, and permitted more efficient management of funds. By the end of the 1980s, automatic tellers for electronic transfer and withdrawal of funds from both demand and savings accounts had become common, and were more numerous than bank branches. Many new variants of demand and savings deposits had been created and the distinction between demand and savings deposits in terms of their liquidity became blurred almost to the point of disappearance, though savings deposits still paid higher interest but also imposed higher charges. Deregulation, innovation and technological change resulted in a failure of the quarterly specification for money demand, whether money was defined narrowly or broadly.

These developments in the 1980s led to the estimated demand functions performing poorly, with unstable money demand and with a highly variable velocity of circulation. The econometric tests also became much more sophisticated than in earlier periods. Among these, econometric tests of the money and income time series showed that they were not stationary. To deal with this, cointegration analysis became one of the preferred techniques and showed that money and income were indeed cointegrated, as were interest rates with them over many periods.

9.2 Common problems in estimation: an introduction

This section is intended to show that the estimation of the money-demand function is not a simple and straightforward matter, and that application of the classical least-squares regression technique to its estimation need not provide reliable estimates. The section provides a brief treatment of the common problems encountered in money demand estimation and is not meant to provide a complete, in-depth or rigorous treatment of the econometric problems discussed or of the appropriate econometric techniques. These are left to econometrics textbooks such as Davidson and Mackinnon (1993).

The general form of the demand function for real balances implied by the transactions, speculative, buffer stock and precautionary analyses is of the type:

$$M^{\mathrm{d}}/P = m^{\mathrm{d}} = m(R_1, \ldots, R_m, \pi^{\mathrm{e}}, y, w) \tag{4}$$

where:

M = nominal balances
m = real-money balances
P = price level
π^{e} = expected rate of inflation
R_i = rate of return on ith near-money asset, $i = 1, \ldots, m$
y = real income/expenditures
w = real wealth.

The following subsections consider some of the econometric issues that arise in the estimation of such a money demand function.

9.2.1 Single equation versus simultaneous equations estimation

From a *general equilibrium* viewpoint, the rate of return on each of the near-money assets is influenced by the demand and supply of money and also by the demands and supplies of risky assets. A general empirical study of the demand for money would then simultaneously estimate the demand and supply functions for all the financial assets, where the demand function for the *i*th asset is:

$$x_i = x_i(R_1, \ldots, R_m, R_{m+1}, \ldots, R_n, \pi^e, y, w) \tag{5}$$

The definitions of the symbols are:

x_i = real quantity of the *i*th monetary asset, $i = 1, \ldots, m$
R_j = rate of return on the *j*th non-monetary asset, $j = m+1, \ldots, n$.

Note that from a rigorous general equilibrium viewpoint, each asset should be homogeneous. A general equilibrium study becomes an extremely large enterprise and poses its own econometric problems. Most studies of the demand for money have been partial and, for statistical and other reasons, have used various degrees of aggregation in defining money. They also often confine the explanatory variables to one rate of interest and either income or expenditures or wealth. However, whether or not one is estimating the demand functions for several assets simultaneously, it is important to consider the cross-equation restrictions that the relevant theory might imply for them. We illustrate these in the following for the case of the allocation of a portfolio between money and bonds, as analyzed in Chapter 5.

9.2.2 Estimation restrictions on the portfolio demand functions for money and bonds

Chapter 5 implied that the general form of the speculative demand functions for assets is:

$$x^d_i = x^d_i(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\rho}, W) \qquad i = 1, \ldots, n-1 \tag{6}$$

where $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\rho}$ are respectively the vectors of the mean returns, the standard deviations and the correlation coefficients among the values of the assets, and W is the wealth allocated among the assets. (6) and the portfolio budget constraint imply that:

$$x_n = W - \Sigma_i\, x^d_i(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\rho}, W) \qquad i = 1, \ldots, n-1 \tag{7}$$

so that the demand function for one of the assets must be derived as a residual from the estimated demand function of the other assets. Alternatively, if the demand functions for all the assets are being estimated, the appropriate cross-equation restriction must be imposed on the estimating equations. As an illustration of this, the restrictions imposed by (7) for the two-asset case of money (M) and the composite bond (B) are set out below.

Suppose the estimating equations for M and B are linear and are specified as:

$$M = a_0 + a_1 R_m + a_2 R_b + a_3 W \tag{8}$$

$$B = b_0 + b_1 R_m + b_2 R_b + b_3 W \tag{9}$$

where R_m is the nominal return on money and R_b is the nominal return on bonds. The budget constraint on M and B is:

$$M + B = W \tag{10}$$

Substituting (8) and (9) into (10) yields:

$$(a_0 + b_0) + (a_1 + b_1)R_m + (a_2 + b_2)R_b + (a_3 + b_3 - 1)W = 0 \tag{11}$$

To satisfy (11) for all possible values of the variables, each term in it has to be zero. Therefore, we must have:

$$a_3 + b_3 = 1 \tag{12}$$

$$a_i + b_i = 0 \qquad i = 0, 1, 2 \tag{13}$$

Failure to impose these restrictions on the estimated coefficients in the simultaneous estimation of both demand functions will generally yield estimated values of the coefficients that are not consistent with the budget constraint and are therefore not valid. In cases where a single demand function, say for money, is estimated, and its estimated coefficients seem to be quite plausible, the *implied* values of the coefficients for the bond equation may not prove to be accurate or even plausible. For example, if the estimated elasticity of the demand for money is much larger than one, this would in turn imply that the elasticity of the demand for all other financial assets is correspondingly less than one, which may not be plausible for the economy and the period in question, thereby leading to a rejection of the estimated money demand function. Therefore, if it is feasible, it would be better to estimate *simultaneously* the complete system of demand equations and impose appropriate restrictions on the coefficients. However, this is not always feasible and often exceeds the researcher's interests, so that most studies tend to confine themselves to the estimation of only the money-demand function.

9.2.3 The potential volatility of the money demand function

Note that the coefficients a_i, $i = 0, 1, 2, 3$ in the money-demand function (8) depend upon the means, the standard deviations and the correlation coefficients of the expected terminal values of the assets, for all of which the subjectively – not objectively – expected future (not the past actual) values are the relevant ones. If these characteristics of assets change, the implied coefficients will change and the demand functions will shift. In the real world, subjective expectations on the returns and future values of the financial assets continuously shift for a variety of reasons, so that the subjectively based characteristics of assets are constantly changing. These sources of shifts can be classified into (i) shifts in subjective probability estimates because of changing market conditions, (ii) shifts in policies which shift the outcomes and their probability functions, and (iii) innovations in the payments mechanism, such as the introduction of ATMs and electronic banking.

Keynes (1936, Ch. 13) focused on (i) and argued that the expectations of asset returns, and hence of these characteristics, are very volatile. This argument implies that the demand functions for money and other financial assets would be constantly shifting, so that they could not be properly estimated or, if estimated, would be worthless – unless the nature of the shift could be specified and adjustments made for it – as guides for future policies.

The Lucas critique (in Chapter 17) of estimated functions used for policy purposes focuses on (ii) above and argues that, if a change in policy – for example, in the monetary regime, tax laws, banking and financial regulations, relevant political stance, etc. – shifted the characteristics of the returns to the assets, the demand functions would shift and the prior estimated forms would no longer be valid. Hence, specific forms of the demand functions will not hold across policy regimes.

The above arguments caution that, since the money demand and supply functions, as well as other relevant policy functions, are constantly changing and definitely do so over decades, the validity of using data over long periods of time to estimate a demand function with constant coefficients should be extremely suspect. This is especially so in a period of financial innovation, which keeps changing the relative characteristics of the existing assets and, over time, keeps adding newer ones to the marketplace.

9.2.4 Multicollinearity

Another statistical problem encountered in partial studies is the *multicollinearity problem*. Suppose that the demand for money is related to both income and wealth but that income and wealth are themselves highly correlated. The estimate of the relationship between money balances demanded and income is then influenced by the relationship between income and wealth and vice versa, so that the estimated relationship may not be an accurate measure of its actual value.

Similarly, the various rates of return are highly correlated, so that the estimates of the coefficients of the rates of return in the money demand function in the economy also tend to be biased and must be treated with caution.

If there is fairly close correlation among a set of variables, one solution to the multicollinearity problem is to use only one of the variables in the set and interpret its estimated coefficient as representing the collective effect of all the variables in the set. For instance, given the close correlation among the interest rates, most money demand functions include among the independent variables only one interest rate in order to avoid multicollinearity. This is usually a short-term rate, such as the Treasury bill rate. However, some studies include both a short-term and a long-term interest rate. As between current income and permanent income or wealth, while some studies include only current income, others include permanent income, with multicollinearity between these two variables preventing the simultaneous inclusion of both.

9.2.5 Serial correlation and cointegration

Most regression techniques assume that the error terms are serially uncorrelated and have a constant variance. These should be checked for the estimated error. If it does not satisfy these conditions, as often proves not to be so, the estimated coefficients will be biased and the appropriate techniques that can ensure unbiased estimates have to be used. The techniques often used for correcting for serial correlation include estimating the money demand function in a first-difference form and using a technique with a built-in correction for the relevant order of serial correlation.

Regression analysis used for deriving the money-demand function assumes that the variables are *stationary*. A variable is not stationary if it has a trend or/and serial correlation. Many of the variables in the money demand function, such as income and the money stock, are not stationary. If this happens, the use of classical regression techniques, such as one-stage

least squares, two-stage least squares, etc., yield biased estimates of the coefficients of the independent variables. The preferred procedure in such cases is that of *cointegration analysis*.

9.3 The relationship between economic theory and cointegration analysis: a primer

This section presents a brief introduction to stationarity and cointegration analyses. The reader is referred to econometric textbooks for a proper treatment of these topics.

9.3.1 Economic theory: equilibrium and the adjustment to equilibrium

An economic theory is intended to explain the determination of the actual values of a selected economic variable or of several economic variables. As the starting point for the following exposition, we focus on the determination of a single economic variable, say y. The theory on its determination examines three questions:

1 Does an equilibrium relationship exist between the dependent variable y and its explanatory variables $\boldsymbol{x}$ $(= x_1, x_2, \ldots)$? Suppose that such a relationship exists and is of the form:

$$y_t = \alpha_0 x_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \tag{14}$$

where x_0 is taken to be constant, and y and $\boldsymbol{x}$ could be the levels of the variables, their first differences or rates of change, etc., or some mix of these, and the relationship could be linear or non-linear. For the following exposition, the equilibrium relationship is assumed to be among the levels (or the log values) of the variables and to be a linear one. The estimation equation, expanded to include lagged values of the dependent variable y and the explanatory variables $\boldsymbol{x}$ on the right-hand side, becomes an autoregressive distributed lag (ARDL) equation, whose treatment is presented in the Appendix to this chapter.

If all the variables in the relationship are stationary, classical least-squares techniques can yield unbiased estimates of the coefficients of the variables. However, if some of the variables in the relationship are not stationary, these techniques do not yield unbiased estimates. The cointegration estimation technique is likely to yield better results. To determine the technique that is appropriate, the stationarity or otherwise of each of the variables has first to be determined by stationarity tests. These are discussed later.

2 Is the equilibrium unique? It is assumed that there is a unique equilibrium for the given structural specification of the equations of the model.

3 Is the equilibrium relationship between y and $\boldsymbol{x}$ stable or unstable? Assuming it to be stable, there would be a dynamic adjustment path during the disequilibrium following a disturbance to the equilibrium relationship. The dynamic path can be of different types, requiring different specifications of the adjustment process.[3] It is often not clear which one is the empirically relevant process for the structure in question. The most common assumption is that the adjustment process is linear (or log-linear). The estimation of the

3 Among these are the partial adjustment models, the error-learning model, etc., presented in Chapter 8.

adjustment process and its implications for the stability of equilibrium are discussed later under the heading of error-correction models (ECMs).

9.4 Stationarity of variables: an introduction

The equilibrium relationship between the endogenous variable y and the vector $\boldsymbol{x}$ of explanatory variables was specified above as:

$$y_t = \alpha_0 x_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \tag{15}$$

Estimation of this equilibrium relationship by classical regression techniques requires that each of the variables be stationary.

A variable z_i is said to be stationary if its mean, variance and covariances with the other variables in the relationship are finite and constant. The usual symbols for these are:

$$E(z_i) = \mu_i$$

$$V(z_i) = \sigma_{ii} = \sigma_i^2$$

$$\text{COV}(z_i, z_j) = \sigma_{ij}$$

The stationarity of z_i implies that these moments of its distribution will remain unchanged, except for random differences, over different sample periods. Conversely, if estimation over different sample periods yields different estimated values of these moments, the variable is likely to be non-stationary. If any of the variables in the relationship implied by the theory is non-stationary, then the estimates obtained by classical least-squares regressions of the equilibrium relationship among the variables will differ among the various sample periods, so that the estimated relationships will not accurately reveal the true relationship.

Causes of non-stationarity

The potential causes of non-stationarity are:

1 The mean value of the variable is not stationary, due to a trend.
2 The variance of the variable and its covariances with other variables are not stationary. This is due to serial correlation.

If the variables in the estimation are not stationary due to serial correlation, two different types of estimation procedures can be attempted for estimating the true equilibrium relationship. One of these is to render the data series stationary prior to estimation, such as by employing a procedure for eliminating serial correlation. To render a series (with serial correlation) stationary, each would be differenced once or more times until its derived series is stationary. Alternatively, a correction for serial correlation, such as the Cochrane–Orcutt method, can be used in the estimation process. Classical regression techniques, such as one-stage or two-stage least squares, often employ such procedures to deal with non-stationary time series.

An alternative to the above procedure is to use the following property of the equilibrium relationship, with y as the dependent variable and, for illustration, only x_1 and x_2 as the

explanatory ones. Assume, as before, that the equilibrium relationship is linear in the levels (or log values) of the variables, so that it is of the form:

$$y_t = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 \tag{16}$$

which can be rewritten as:

$$y_t - \alpha_0 - \alpha_1 x_1 - \alpha_2 x_2 = 0 \tag{17}$$

In this equation, since the right-hand side (which is zero) is stationary, the composite variable $(y_t - \alpha_0 - \alpha_1 x_1 - \alpha_2 x_2)$ on the left-hand side must also be stationary. Hence, while the individual variables are not stationary, their linear combination given by (17) would be stationary. The appropriate linear combination is one with the coefficients $(1 - \alpha_0 - \alpha_1 - \alpha_2)$. Note that α_0 is the (coefficient of the) constant term. The vector $(1 - \alpha_0 - \alpha_1 - \alpha_2)$ is called the cointegrating vector (in this case, with the coefficient of the dependent variable normalized to unity) and (17) is called the *cointegrating equation*. Empirical analysis requires an appropriate estimation procedure that will provide unbiased estimates of this vector. A few points about the above relationship need to be noted.

- Multiplying each of the coefficients by a constant yields a stationary variable, so that any multiple of the cointegrating vector is also a cointegrating vector.
- It is quite appropriate to set the coefficient of the endogenous variable as unity, so that it is customary to normalize the cointegrating vector in this way.
- The signs of the coefficients of the explanatory variables in the cointegrating vector and equation are the reverse of those in the equilibrium relationship.
- If the equilibrium relationship is linear in the logs of the variables, then the cointegrating vector will specify (with signs reversed) the elasticities of the endogenous variable with respect to the explanatory variables.

The next section discusses the sources of non-stationarity of the variables and the estimation procedures for determining whether a variable is stationary or not.

A non-stationary mean due to a trend

A trend in a variable will make its mean non-stationary. An example of this occurs if:

$$z_t = \alpha_0 + \alpha_1 t + \mu_t \tag{18}$$

where z is the variable in question, t is time and μ is white noise.[4] In this case, data samples over different periods will yield different mean values of z, since:

$$Ez_t = \alpha_0 + \alpha_1 Et$$

A variable that is nonstationary because of the presence of a trend can be transformed by removing the trend into a corresponding variable (i.e. $z_t - \alpha_1 t$) that is stationary. Such a variable is said to be trend-stationary (TS).

4 That is, μ is a stationary disturbance term, with a zero mean and finite and constant variance and covariances.

Non-stationary variances and covariances because of serial correlation

This type of non-stationarity arises if the variable behaves according to:

$$z_t = \alpha_0 + z_{t-1} + \mu_t \tag{19}$$[5]

where μ_t is white noise. z_t is said to follow a random walk if the constant term α_0 is zero; it follows a random walk with drift if α_0 is not zero. The value of z_t depends on the actual value of z_{t-1} (which includes the actual value of μ_{t-1}), so that there is a stochastic tendency for the mean of the variable to change over different data samples.

Rewrite this equation as:

$$\Delta z_t = z_t - z_{t-1} = \alpha_0 + \mu_t \tag{20}$$

where $\Delta z_t (= z_t - z_{t-1})$ is stationary. Therefore, a variable that is non-stationary because it follows a random walk can be rendered stationary by taking its first difference. Such a variable is called difference-stationary (DS). If taking the first difference of a series makes it stationary, it is said to be integrated of order 1, which is written as I(1).

Note that if a series is I(1), taking its first difference will yield a stationary series. But if the series also has a time trend, the first difference of the series will still possess a trend, so that it will not be trend-stationary. An adjustment for this trend will have to be made in the estimation procedure.

Non-stationarity because of a shift in the value of the variable

Note that a variable may be stationary but that its data sample may indicate non-stationarity because of a shift at some point in its time series. In this case, classical least-squares can still be used with the shift captured through the use of a dummy variable.

9.4.1 Order of integration

In the general case of serial correlation, a variable may follow the process:

$$z_t = \alpha_0 + z_{t-p} + \mu_t \tag{21}$$

In this case, the variable would have to be differenced p times to arrive at a stationary series. The variable is then said to be integrated of order p and is designated as I(p).

In the case of difference-stationary data, while using the appropriate number of differences of the variables does eliminate the problems posed by the non-stationarity of the levels of the variable, a regression using differenced data eliminates the relationship among the levels of the variables, so that the regression will not provide estimates of the long-run relationship between the *levels* of the dependent and the independent variables in the estimating equation. Therefore, the use of differenced data is not a proper strategy for finding the equilibrium relationship among the levels of the variables. For example, in the context

5 This equation implies serial correlation of the error over time since substitution in it of the equation for the lagged term implies that $z_t = 2\alpha_0 + z_{t-2} + \mu_t + \mu_{t-1}$, so that the error $(\mu_t + \mu_{t-1})$ in period t will be correlated with the error μ_{t-1} in period $t-1$. Repeated substitutions of the equation for the lagged term will, in fact, yield an equation whose error term is correlated with the errors in all previous periods.

of the money-demand function, the underlying theory implies an equilibrium relationship between the levels of the variables, so that using differenced data will not provide an estimate of this function.

Note that if a series is I(p), taking its pth difference will yield a stationary series. But if the series also has a time trend, the pth difference of the series will still possess a trend, so that it will not be trend-stationary and an adjustment for this trend will have to be made in the estimation procedure.

9.4.2 Testing for non-stationarity

Since non-stationarity can arise from both a trend and serial correlation, the appropriate test for stationarity has to simultaneously test for both these. The following discusses such tests.

Suppose that the variable z follows an autoregressive, non-stationary, data-generating process with a one-period lag:

$$z_t = a_0 + a_1 t + a_2 z_{t-1} + \mu_t \tag{22}$$

where t is time and μ_t follows a stationary process. Subtracting z_{t-1} from both sides,

$$\Delta z_t = a_0 + a_1 t + (a_2 - 1) z_{t-1} + \mu_t \tag{23}$$

If $a_2 = 1$, z_t is I(1). The test for $a_2 = 1$ as against $a_2 < 1$ is called a unit root test. Such a test is referred to as the Dickey–Fuller (DF) unit root test. The estimation of this equation can yield the following results:

1 If $\hat{a}_0 = \hat{a}_1 = 0$ and $\hat{a}_2 = 1$, then z follows a random walk and its series is I(1).
2 If $\hat{a}_0 \neq 0, \hat{a}_1 = 0$ and $\hat{a}_2 = 1$, then z follows a random walk with drift and its series is still I(1).
3 If $\hat{a}_2 = 1$ and $\hat{a}_1 \neq 0$, then z has a trend and is trend-stationary.

A more sophisticated test for the sources of non-stationarity is provided by the Augmented Dickey–Fuller (ADF) unit root test.[6] This test allows for higher-order autoregressive processes and is based on the estimation of the equation:

$$\Delta z_t = a_0 + a_1 t + (a_2 - 1) z_{t-1} + \sum_{j=1}^{n} b_{ij} \Delta z_{t-j} + \mu_t \tag{24}$$

which allows for the impact of n lagged values of the variable. The ADF unit root test is for the null hypothesis that $a_2 = 1$, against the alternative that $a_2 < 1$.[7] Failure to reject the null hypothesis implies non-stationarity of the series.

If the ADF and other tests[8] for the data series of the variables in a relationship show that at least some of the series are I(p), $p \geq 1$, the relationship has non-stationary variables so that,

6 It is also necessary to supplement the ADF tests with other tests of non–stationarity, such as the Phillips–Perron test.

7 Note that a structural break in the data can sometimes be mistaken for a unit root, so that the appropriate checks and corrections for this possibility are needed.

8 Other tests for the non–stationarity of a series include the Phillips–Perron test, the Durbin–Watson test, etc.

as mentioned above, the classical regression techniques – such as ordinary least squares – will not provide unbiased and consistent estimates of the coefficients of the relationship. An appropriate technique would be cointegration.

9.5 Cointegration and error correction: an introduction

The cointegration technique is based on the assumption of an equilibrium (linear or log-linear) relationship among the variables, which implies that two or more variables that are individually non-stationary but are integrated of the same order possess a linear combination of a one-degree lower order of integration.[9] Therefore, if all the variables are I(1) and are cointegrated, then their cointegrating equation would yield a composite variable of order I(0), i.e. it would be stationary. As explained earlier in the discussion on the connection between an equilibrium relationship and cointegration, if the equilibrium relation among a set of I(1) variables is linear (log-linear), the existence of such a linear (log-linear) combination is the equilibrium relationship implied by the relevant theory. Cointegration techniques attempt to estimate whether such a combination exists and, if so, what is the cointegration vector.[10] The cointegration equation based on such a vector is then treated as an estimate of the long-run equilibrium relationship.

If the variables are all I(p), then their cointegrating vector, if it exists, will yield a variable which is I($p-1$).

In practice, problems in using cointegration analysis arise if the variables in the relationship implied by the theory are of different orders of integration. If y is I(2) and some of the x_i, $i = 1, 2, \ldots, n$, are I(1) while others are I(2), the successful[11] application of the cointegration technique to the I(2) variables only would yield a cointegrating equation that provides an I(1) composite variable. The I(1) estimate of this composite variable can then be used along with the I(1) variables in the error-correction estimation, discussed later. A similar procedure would have to be used if y is I(1) and some of the x_i variables in the relationship implied by the theory are I(0) while others are I(1).

If the dependent variable is of a lower order of integration than some or all of the explanatory variables implied by the theory, then it is inappropriate to use cointegration analysis. This would occur if y is I(0) (i.e. stationary) while some or all the explanatory variables are I(p), $p \geq 1$.

Estimation problems therefore arise if the variables are of different orders of integration. In such a case, it might be more appropriate to use a cointegration procedure that allows such variability. Pesaran *et al*. (2001) provide such a procedure.

Relationship between cointegration results and economic theory

Let the relationship derived from economic theory be of the form:

$$y_t = \alpha_0 x_0 + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \tag{25}$$

9 This requires the assumption that the underlying equilibrium relationship among the variables is linear (or log-linear).

10 Note that not all data samples may show this relationship.

11 This really requires an equilibrium relationship among the I(2) variables only, which may not be what the theory asserts.

If the variables $y, x_1, \ldots, x_n$ are all I(1), the general form of the cointegrating vector, if one is found, is:

$$f(y, x_1, \ldots, x_n) = 0$$

This relationship is log-linear or linear depending upon whether the data was in logs or not. The form of the *cointegrating equation* is:

$$y_t - \alpha_0 x_0 - \alpha_1 x_1 - \alpha_2 x_2 - \cdots - \alpha_n x_n = 0 \tag{26}$$

where x_0 represents the constant term. As mentioned earlier, its signs of the coefficients of the explanatory variables have to be reversed to arrive at the original equation (25).

Because of potential econometric problems, various econometric checks (discussed later) are applied to check the econometric acceptability of the estimated coefficients. However, even if the estimate is acceptable on the basis of the econometric tests, the estimated cointegrating vector may still not be a plausible estimate of the true equilibrium economic relationship. From the perspective of economic theory, this plausibility is judged by checking whether the signs of the cointegrating vector are consistent with those implied by the theory and whether the estimated magnitudes of the normalized cointegrating equation are plausible in terms of the theory, intuition and estimates obtained by other studies. If this is not so, the estimated cointegrating vector will have to be rejected as an estimate of the true equilibrium relationship.

Deviations from the equilibrium relationship: the error-correction assumption for adjustments in disequilibrium

It was assumed earlier that the equilibrium relationship between the endogenous variable and the vector of explanatory variables is stable and unique. Cointegration literature labels this equilibrium relationship – and its estimate by the cointegrating vector – the *long-run* relationship.

Since the equilibrium has been assumed to be stable, any deviations from it will be corrected through an adjustment process.[12] In cointegration analysis, this adjustment process is often referred to as the *dynamic adjustment* or as the *short-run relationship* between the endogenous variable and the explanatory variables. Cointegration techniques assume that the dynamic adjustment follows a linear or log-linear error-correction process, rather than some other one. This process is in the nature of a partial linear adjustment each period.[13]

The cointegration literature refers to its adjustment estimation technique as "the *error-correction model*" (ECM). This model specifies the change in the endogenous variable y

12 Engle and Granger (1987) present a theorem showing that cointegration among a set of variables implies short-run dynamics that return the variables to the long-run cointegrating relationship. This follows from the assumption of the stability of equilibrium and that the cointegrating vector captures the equilibrium relationship.

13 This error-correction element is somewhat similar to the linear first-order partial adjustment process, as well as to the error-learning process, both explained in Chapter 8. For comparison, the first-order PAM for the adjustment in y would be $y_t - y_{t-1} = \gamma(y_t^* - y_{t-1})$, $0 \geq \gamma \geq 1$, where y_t^* would be the long-run or desired value of y_t.

as a function of last period's error between the actual and the equilibrium value of the dependent variable and of the change in each of the explanatory variables of the equilibrium relationship. Other variables, provided that they are stationary, can also be introduced in the ECM. The linear specification of the error-correction element of the ECM is specified as:

$$\Delta y_t = \theta(y_{t-1} - y^*_{t-1}) + \cdots \tag{27}$$

where $\Delta y_t = y_t - y_{t-1}, y^*$ is the equilibrium value (calculated from the estimated cointegrating vector), and y changes each period by the fraction θ of the previous period's deviation of the actual value from the equilibrium value y^*. For the equilibrium to be stable (equilibrium-reverting), we need $\theta \leq 1$. The complete ECM equation would have the form:

$$\Delta y_t = a_0 - \sum_{i=2}^{p} a_i \Delta y_{t-i} + \sum_{j=1}^{n} \sum_{i=1}^{q} b_{j,t-i} \Delta x_{j,t-i} - \theta \mathrm{ECM}_{t-1} + \eta_t \tag{28}$$

where the lag lengths p and q have been optimally determined and:

$$\mathrm{ECM}_{t-1} = y_{t-1} - \hat{\alpha}_0 - \sum_{j=1}^{n} \hat{\alpha}_j \boldsymbol{x}_{jt-1} \tag{29}$$

9.5.1 Cointegration techniques

The two popular cointegration procedures for determining the equilibrium relationship or relationships among non-stationary variables are the Engle–Granger (Engle and Granger, 1987) and the Johansen – also called the Juselius-Johansen – procedures (Johansen and Juselius, 1990; Johansen, 1988, 1991). The most common application of these procedures is when all the variables are I(1).

Engle–Granger method for a reduced-form equation[14]

For the estimation of the cointegration vector and its associated error-correction dynamic adjustment equation, the Engle–Granger method uses a two-stage procedure. In the first stage, it estimates the cointegrating vector among the I(1) variables for a given equilibrium relationship and tests the residuals for stationarity. If these residuals are stationary, as they should be if all the variables are I(1), the second stage uses them to estimate the dynamic short-run response of the dependent variable by the error-correction model.

The Engle–Granger technique is quite appropriate if all the explanatory variables are exogenous. Often, a model has several endogenous variables, so that it possesses several equilibrium relationships among its variables. In this case, the Johansen procedure would be preferable to the Engle–Granger one.

14 A reduced-form equation has only one endogenous variable, whose explanatory variables are all exogenous.

Johansen cointegration procedure for a model with several endogenous variables[15]

In a model where more than one variable is endogenous, there would be more than one equilibrium relationship among the variables. The Johansen cointegration procedure is then the preferable one since it treats all the variables in the estimation process as endogenous and tries to simultaneously determine the equilibrium relationships among them. In addition, this procedure provides estimates of the cointegrating vectors and the error-correction model in one step. These advantages have made the Johansen procedure the more common one in the cointegration literature.

Assuming that all the variables being considered are I(1), the Johansen procedure (a) takes all the I(1) variables to be as if endogenous and related by a vector-autoregressive (VAR) structural model, (b) uses the maximum likelihood estimation for the VAR model, and (c) derives a set of cointegrating vectors. The number of cointegrating vectors is determined by the eigenvalue and trace tests. The maximum number of independent equilibrium relationships that can exist among a set of endogenous variables has to be one less than the number of variables.[16] Therefore, the maximum number of significant cointegrating vectors should be one less than the number of variables in the VAR model.

The propensity of the Johansen procedure to yield several (significant) cointegrating vectors among the variables is an asset but also raises two troublesome issues:

1 Which vector should be treated as the estimate of which one of the equilibrium relationships among the variables? That is, a choice has to be made among the available cointegration vectors for the particular economic relationship being sought. This choice is usually made on the basis of the signs implied by the theory for the coefficients and the estimated magnitudes of the coefficients falling within a plausible range.
2 Any linear combination of the estimated cointegrating vectors is also an admissible cointegrating vector. Therefore, one can generate an infinite number of combinations, many of which are usually likely to fit the requirements of the appropriate signs and magnitudes being sought for a specific relationship. The linear combinations can be searched for this purpose. However, this search can easily degenerate into "vector-mining."

To illustrate, in some applications of the Johansen technique to money-demand estimation, it is found that the elements of none of the cointegration vectors possess signs consistent with the a priori expectations on the elasticities of the money demand function. Alternatively, these elements could be such as to imply implausible magnitudes of the elasticities. These problems could arise from the limited sample size, inaccuracies in the data, misspecification in the set of variables, breaks in the data, etc. But it is also possible to argue that, since a linear combination of the cointegrating vectors is also a cointegrating vector, one could try to find that linear combination of the cointegrating vectors such that the elements have the desired signs and magnitudes in a plausible range. However, this amounts to "mining the vectors," so that the results often fail to convince other researchers.

15 Such a model would be a structural one.

16 Given the assumption of linearity (log-linearity) of the equilibrium relationships among n variables, there can be at most $(n-1)$ independent relationships among them.

To conclude, while the Johansen technique provides econometric evidence on the existence of long-run relationships among a set of variables, the identification or derivation of the structural coefficients of the model from the elements of the cointegration vectors can be quite problematical.

9.6 Cointegration, ECM and macroeconomic theory

Economic theory often implies more than one long-run relationship among any given set of economic variables. For example, in the IS–LM model, money demand depends upon national income and interest rates, while national income – as do interest rates – depends on the money supply, which equals money demand in equilibrium. Assuming these three variables to be all I(1), such a simultaneous determination of economic variables implies the possible existence of a maximum of two cointegrating vectors among them. In general, for n variables, there could be $(n-1)$ independent cointegrating vectors. This poses a problem since the cointegration technique does not identify a given cointegrating vector with a specific economic relationship. For instance, suppose two cointegrating vectors are found among money, income and interest rates. The econometric estimation by itself does not make it clear which one of the cointegrating vectors specifies the money demand relationship. This has to be decided by the researcher on the basis of the signs imposed by economic theory on the coefficients of the money demand relationship and on the basis of the plausibility of the magnitudes of the elements in the cointegrating vectors. The elements of the selected cointegrating vector are then taken to specify the respective long-run coefficients of the linear (or log-linear) money demand function.

Now, assuming that a cointegrating vector exists, the ECM can be used to capture the adjustment of the dependent variable to the long-run equilibrium specified by the cointegrating vector. Among the characteristics of the ECM are:

1 It defines the deviation from the long-run value as the "error" and measures it by the residual, i.e. the difference between the actual value of the dependent variable and its estimated value based on the selected cointegrating vector.
2 It specifies the first difference of the dependent variable as a function of this error lagged one period, the I(0) variables and the first differences of the independent I(1) variables.[17] Appropriate lags in the latter are introduced at this stage.
3 The coefficient of the lagged residual is the error-correction coefficient and specifies the speed of adjustment of the dependent variable to its long-run value.
4 The estimated coefficients measure the short-run movements in the dependent variable in response to fluctuations in the independent variables.

9.7 Application of the cointegration–ECM technique to money demand estimation

To illustrate the application of the cointegration–ECM procedure to money demand, let the long-run money demand function be:

$$m^{\mathrm{d}}{}_t = a_0 + a_R R_t + a_y y_t \tag{30}$$

17 This model is valid only if the estimated error is stationary.

Assume that the data series for m, R and y are all I(1). Let their estimated cointegrating *vector* be $(1 - \hat{a}_0 - \hat{a}_R - \hat{a}_y)$ in which the second, third and fourth elements have the opposite sign to that of the respective coefficient on the right-hand side of the equation. Let the estimated value of m^{d} from this cointegrating vector be $\hat{m}^{\mathrm{d}}$. That is,

$$\hat{m}^{\mathrm{d}}_t = \hat{a}_0 + \hat{a}_R R_t + \hat{a}_y y_t$$

The error-correction model is then specified as:

$$\Delta m^{\mathrm{d}}_t = \alpha z_t + \beta(m^{\mathrm{d}}_{t-1} - \hat{m}^{\mathrm{d}}_{t-1}) + \gamma \Delta x_t + \eta_t \tag{31}$$

where $(m^{\mathrm{d}}_{t-1} - \hat{m}^{\mathrm{d}}_{t-1})$ is the *lagged* error and z is a vector which includes the constant term and any I(0) variables. Since x is the vector of the independent variables which are I(1) and included in the cointegrating vector, $\Delta x_t (= x_t - x_{t-1})$ is I(0). Under our assumptions on the money demand function, x would include R and y, which were assumed to be I(1) and are in the theoretical specification of the demand function. Since there are no other independent variables in this function, z would consist only of the constant term.

But if only m and y were I(1) while R was I(0), the cointegration would be appropriate only over m and y. If the estimated cointegrating vector met the theoretical restrictions for the money-demand function and therefore was accepted as the long-run money demand function, the error-correction equation (31) would specify z by a constant term and R, while x would be specified by the single variable y.[18]

To conclude this section, given that the data on the money stock and income – and possibly on other variables in the money demand function – are almost always at least I(1), it is inappropriate to use the standard least-squares regression methods. This has led to the popularity of the cointegration–ECM procedure for the estimation of money demand functions. An appealing feature of this procedure is the separation of the long-run money demand function from its dynamic short-run form in a simultaneous econometric estimation of the two. One defect of the Johansen procedure arises if one or more of the variables are I(0) but have a structural break, which makes their series appear to be I(1).[19]

9.8 Some cointegration studies of the money-demand function

We examine a few studies that used the cointegration–ECM for their findings. Among these, Baba *et al*. (1992) considered the standard money-demand equation, with only the interest rate and income as the explanatory variables to be misspecified for several reasons. They claimed that these variables suffer from the omission of the inflation rate, inadequate inclusion of

18 This procedure poses a conundrum. Monetary theory implies that the level of money demand is a function of the level of the interest rate. That is, $m^{\mathrm{d}} = m^{\mathrm{d}}(R, y)$. However, what the ECM in this procedure yields is a relationship between Δm^{d} and R, so that this procedure does not yield the relationship asserted by the theory. An alternative would be to run a least-squares regression between the money demand calculated from the estimated cointegrating vector and the interest rate. This would be a two-stage procedure for estimation of the long-run money demand function, followed by estimation of the ECM.

19 The bounds-testing cointegration procedure (Pesaran *et al*, 2001) may be preferable in such a case since it does not require a priori knowledge on the variables being I(0) or I(1).

the yield on money itself, inadequate adjustment for financial innovation in the yields on alternative assets, exclusion of the risk and yield on long-term assets and, finally, improper dynamic specification. On the last item, they considered the partial adjustment model or the usual corrections made for serial correlation, such as the Cochrane–Orcutt technique, to be unacceptable for various reasons.

Baba *et al*. therefore estimated a more elaborate M1 demand function using the cointegration–ECM technique for the USA for 1960–88. They reported finding a stable cointegrating M1 demand function consistent with theory. Further, their finding was that the short-run money demand dynamics were adequately captured by the error-correction specification. They found a significant impact of inflation, apart from those of interest rates, on M1 demand. The inclusion of a long-term bond yield, adjusted for risk, was also significant and important for explaining the changes in velocity. However, their variable for the yield on alternative assets was a construct which included adjustments for the changing availability of financial instruments and the time required in the learning process for these instruments to be fully adopted. They concluded that if the yield data is not suitably adjusted for these factors, the mere inclusion in the estimated equations of the own-interest rates on financial assets will lead to the rejection of parameter constancy and stability.

These findings of the Baba *et al*. study point to the usefulness of the cointegration–ECM technique and the need to specify properly the variables in the money-demand function. They also stressed that financial innovation had been significant. This leads to instability of the estimated function unless the financial innovation and its pace are properly captured by the data. Unfortunately, the method that works best for capturing this in one study for a given country and given period does not often do equally well over other periods or for other countries, so that the methods for capturing innovations remain varied and somewhat eclectic.

Miller (1991) used the demand for nominal money balances as a function of real income, the nominal interest rate and the price level. His specification for money included M1, M1A, M2 and M3. The alternatives used for the interest rate were the four to six-month commercial paper rate and the dividend/price ratio. The Engle–Granger cointegration–ECM technique was used on the US quarterly data for 1959–87. Of the various monetary aggregates, only M2 was cointegrated with the other variables; none of the other ones were cointegrated.

Hafer and Jansen (1991) used the Johansen procedure for US quarterly data for 1915–88 and for 1953–88. In one part of their study, their variables were real money balances, real income and the commercial paper rate, which is a short-term interest rate. They found a cointegrating vector for M1 for 1915–88, though not for 1953–88, and found such vectors for M2 for both periods. For M1 for 1915–88, the long-run income elasticity was 0.89 and the long-run interest-rate elasticity was –0.36. For M2, the former was a plausible 1.08 for 1915–88 and a plausible 1.06 for 1953–88. The long-run interest-rate elasticity for M2 was –0.12 for 1915–88 and –0.03 for 1953–88, with both estimates being statistically significant. These estimates, especially the latter one, are much lower than the corresponding estimated elasticities in the range –0.15 to –0.5 in many other studies.

When Hafer and Jansen replaced the commercial paper rate by the corporate bond rate – a long-term rate – there was still no cointegrating vector for M1 for 1953–88. There was also none for M2 for 1915–88, but there was one for 1953–88. The income elasticity for the latter was 1.13 and the interest rate was –0.09. Overall, the authors

concluded in favor of using M2 over M1 in a long-term relationship with income and interest rates.

Among other studies, Miyao (1996) used M2 for his money variable and estimated a variety of linear functions involving income, an interest rate and the price level. His sample periods for US quarterly data were 1959–88, 1959–90 and 1959–93. For the earlier periods, there were mixed results suggesting both cointegration and no cointegration, while there was no cointegrating vector at all for 1959–93. The author concluded that there were shifts in the data structure in the 1990s, so that an error-correction model was not appropriate for that decade. Further, his conclusion was that a stationary relationship between M2 and output disappeared in the 1990s, so that M2 was no longer a reliable indicator or target for policy purposes.

As pointed out earlier, innovations have shifted the money-demand function over time. Further, cointegration analysis requires long runs of data. To accommodate these, Haug (2006) uses cointegration techniques with unknown shift points to study the demand for M0, M1, M2 and related money measures for Canada covering several periods, the longest one being 1972–97. Among other criteria for acceptance of findings, Haug uses cointegration rank stability. This study also introduces variables (such as the ratio of currency to the money supply, velocity, and per capita permanent income) that reflect institutional and structural change. The findings, using the long-term interest rate, do show one cointegrating vector for the demand for M1, irrespective of the data time span.[20]

As against studies using M1 or M2 as the preferred monetary aggregate for the Canada and USA, the European Central Bank uses M3 as its preferred monetary aggregate. Coenen and Vega (2001) use cointegration and error-correction analysis to estimate the demand for M3 for the Euro area for the period 1980:Q4 to 1998:Q4. They find a stable long-run demand function for real M3. Their explanatory variables included, in addition to real GDP, short-term and long-term interest rates and the inflation rate. Their estimated long-run income elasticity is 1.13, which they interpret as incorporating wealth effects on money demand.

These differing results clearly indicate that the evidence for recent decades on the cointegration of the variables in the money demand function is not unambiguous or robust for the United States. Similar findings have been reported for the UK (see Cuthbertson, 1991, for a review of some of these) and Canada. While the existence of such a vector cannot be rejected for some form of the monetary aggregate and for some definitions of the independent variables, such a finding is dependent on particular definitions, particular periods and particular cointegration techniques (for instance, see Haug, 2006). Part of the reason for the conflicting findings is the sensitivity of the Johansen cointegrating procedures to the sample size and its poor finite sample properties. But, from the perspective of economic theory, the problem can also stem from numerous shifts in the money demand function due to innovations of various types in recent decades. These shifts imply that there is no stable long-run money-demand relationship over this period. Therefore, the cointegration techniques will not yield the appropriate cointegrating vector, unless the impact of the innovations is somehow first adequately captured in the measurement of the variables, as in the Baba, Hendry and Starr study cited above, and perhaps not even then, since the

20 The Johansen cointegration technique gave more cointegration vectors than the cointegration technique, allowing for unknown shifts for the post-1945 period. For this period, the demand for M0 and the short-term interest rate also gave a cointegrating vector.

innovations have been of numerous types and their collective combination has itself been changing.

9.9 Causality

Since the ECM incorporates lags of the explanatory and other exogenous variables on the right-hand side, its estimates are often used to determine the direction of Granger causality. The criteria for judging one-way versus two-way Granger causality were specified in Chapter 7.

9.10 An illustration: money demand elasticities in a period of innovation

Table 9.1 provides an illustration of the estimates of money demand with a lagged dependent variable and is based on Goldfeld and Sichel (1990). Part of this table is based on Fair (1987), who presented the estimates of money demand for 27 countries.

Income elasticities in Table 9.1

In Table 9.1, the coefficient of the income variable y is the impact elasticity for the quarter and lies in the range 0.039 to 0.118. The long-run elasticity is obtained by dividing the impact elasticity by one minus the coefficient of the lagged dependent variable m_{-1}. The computation of long-run elasticity becomes extremely sensitive to small changes as this coefficient approaches one. In fact, if this coefficient is one or over one, the partial adjustment model leads to a misspecification in its adjustment mechanism. This is clearly so for the USA for 1952:3–1979:3, and almost so for 1974:2–1986:4. The estimates for these periods therefore cannot be relied upon, as a look at the long-run income and interest-rate elasticities clearly shows.

Table 9.1 Estimates of money demand

Country	*Sample period*	y	R_1	R_2	π	m_{-1}	*Long-run elasticities*	
							Income	*Interest*[c]
USA[a]	1952:3–1974:1	0.131	−0.016	−0.030	−0.771	0.788	0.62	−0.075
	1952:3–1979:3	0.039	−0.013	−0.002	−0.889	1.007	−5.57	1.857
	1974:2–1986:4	0.044	−0.018	0.100	−0.823	0.997	14.67	−6
Canada[b]	1962:1–1985:4	0.071	−0.004		−1.66	0.94	1.18	−0.067
UK[b]	1958:1–1986:1	0.118	−0.005		−0.69	0.44	0.21	−0.009

Source: Goldfeld and Sichel (1990), Tables 8.1 and 8.5, of which Table 8.5 is based on Fair (1987).

Notes

a All variables are in logs, except for the inflation rate $\pi(=\ln(P_t/P_{t-1}))$. The dependent variable — is real money balances m, measured by the real value of M1, and y is real GNP. R_1 is the commercial paper rate and R_2 is the passbook savings rate at commercial banks.

b All variables are in logs except the interest rate R_1 which is in levels. The dependent variable m is real balances per capita and the scale variable is income per capita. r_1 is a short-term rate. The reported estimates are taken from Goldfeld and Sichel (1990), Table 8.5, calculated by them from Fair (1987).

c Based on the coefficient of R_1.

Further, only two of the long-run income elasticities in Table 9.1 are plausible. These are 0.62 for the USA for 1952:3–1974:1 and 1.18 for Canada. The estimates of this elasticity for the other two periods for the USA are implausible and, as already argued in the preceding paragraph, the estimated equation as a whole for these periods is highly suspect. Further, Goldfeld and Sichel show that the estimates perform well in simulations only for the first period and that the money demand function shifts sufficiently after 1974 to lead to a breakdown of its estimation in the conventional form used for this table.

Interest-rate elasticities in Table 9.1

From column 4 of Table 9.1, the impact (first quarter) interest-rate elasticities are −0.004 for Canada, −0.005 for UK and −0.016 for the first period for the USA, ignoring the latter two periods for this country. The corresponding long-run interest elasticities are −0.066 for Canada, −0.009 for UK and −0.075 for the USA. For comparison, for Canada for 1956:1–1978:4, Poloz (1980) had reported for M1 the impact and long-run interest rate elasticities of −0.054 and −0.18 respectively. His estimates of the corresponding income elasticities were 0.22 and 0.73. These are somewhat different from those reported in Table 9.1 for Canada, and indicate that one should think in terms of the plausible ranges rather than precise magnitudes for elasticities.

This table also shows significant impact elasticities with respect to the inflation rate, which are in fact higher than the interest-rate elasticities. Since the coefficient of the lagged dependent variable equals $(1 - \lambda)$, the adjustment during the first quarter was only 0.212 for the USA for the first period, 0.06 for Canada and 0.56 for the UK. We have already commented on the instability of the money demand function in the latter two periods for the USA. Fair found instability for 13 out of the 17 countries in his sample.

As discussed in Chapter 4, the Baumol–Tobin inventory model of transactions money demand implies that, at relatively low interest rates relative to brokerage costs, it may not be optimal for economic agents to hold bonds for transactions purposes, whereas doing so would become optimal at higher interest rates, so that the interest elasticity of the transactions demand would vary between −1/2 and 1. Therefore, as Mulligan and Sala-i-Martin (2000) argued, the interest elasticity would be non-linear. Their findings confirm that the interest elasticity of money demand is low at low interest rates.

9.11 Innovations and the search for a stable money-demand function

Financial innovation is a frequent occurrence in the economy. Some types of innovation change the liquidity characteristics of the existing assets or represent the creation of new assets. Other types of innovation are in the payments and banking technologies. Some of the innovations could also be due to the attempt of the financial industry to get around financial regulations. Another is the introduction of new techniques of financial management by firms, households and financial institutions. All of these have occurred during the last three decades, probably collectively at a faster pace than in earlier decades.

Among the new types of assets, in the USA, interest-bearing checking accounts were first introduced as NOW (negotiable orders of withdrawal) and then as super-NOW accounts in the late 1970s and early 1980s. Commercial banks began to issue small certificates of deposit in the 1960s and money-market mutual funds in the late 1970s. These were outside the traditional definition of M1. In the UK, commercial banks and building societies

introduced checkable interest-bearing accounts in the 1980s. In each case, there was a learning period for the public and shifts in the money demand function were evident over many years.

If the innovations merely change the constant term or the coefficients of the independent variables in the money demand function, they can be relatively easy to capture in estimation through period splitting or the use of dummy constant and interactive variables. However, some of the resulting shifts of the money demand function are much more difficult to capture or cannot be captured, and the researcher ends up with the judgment that the money demand function has become unstable.

The desperate search for a stable money demand function

The last three decades have seen a remarkable number of innovations in the monetary sphere. These have resulted in a breakdown of the estimated money demand functions and a large number of innovations by researchers in their estimating equations and techniques. The attempts to find a stable demand function have included changes in the monetary aggregate used as the dependent variable (M1, M2, M3, or their Divisia counterparts). Other attempts have centered around variations in the arguments of the function. These included the use of current income, permanent income, wage income or property income, etc., for the scale variable, and the use of short interest rates, long interest rates, the rate of inflation or a composite index of interest rates, etc., for the interest rate variable.

Still other attempts changed the form of the estimating equation from linear to log-linear and semi-log-linear, or switched to non-linear functions or tried ones with stochastic coefficients, or used transcendental functions. Some other attempts focused on the proper specification of the dynamic adjustment of the actual to desired money balances. The econometric techniques have included the classical regression techniques and cointegration–error-correction models, among others.

This prolific variety of attempts and deviations from the standard money demand equation almost gives one the impression of a field dominated by data mining and the *ad hoc* constructions of a profession desperate to find a stable money demand function to back its theory. While this may sound a rather harsh assessment, it does serve as a reminder of the severe difficulties in finding a stable money demand function during the ongoing innovations of the recent decades.

For the USA, there appears to have been a downward shift in the demand function during the 1970s and an upward shift during the 1980s. In these decades, as in the 1990s, actual money holdings deviated remarkably from the predictions of most estimated money demand models. In terms of velocity, the velocity of M1 increased in the 1970s and decreased in the 1980s in a manner not predicted by these models.

Conclusions

Empirical findings generally confirm the homogeneity of degree zero of the demand for real balances with respect to the price level – and the consequent homogeneity of degree one of the demand for nominal balances – as discussed in Chapter 3. The income elasticity of real M1 with respect to real income has been established as being less than one, even in the long run, though some studies show the income elasticity for real M2 to be even slightly larger than unity. The latter is particularly so for developing economies, in which the

bond and stock markets are not well developed, so that increases in savings are mostly held in savings deposits. Real balances do depend on interest rates, with a short-term rate being usually used in the estimation of M1 demand and a longer term one being used for the estimation of M2 demand. The estimated interest-rate elasticities usually fall in the range from -0.15 to -0.50. In the LDCs, the rate of inflation typically performs better in estimation than the rate of interest and is often used in lieu of the latter, with somewhat similar elasticities. While currency substitution is a theoretical possibility and some studies do confirm its existence for their data sets, empirical studies have not always found it to be so significant that the elimination of the return on foreign currencies from the money demand function leads to much worse results. Most money demand functions are, therefore, estimated without this variable. Not much support has been found for the liquidity trap and it is now hardly ever investigated or even mentioned in empirical studies.

The velocity of circulation of M1 is not a constant in either the short or the long run. In the short run, its annual variation is quite significant even in stable economies without political and economic panics. It is about 3 percent to 4 percent for the USA, but can be much higher in less stable economies. Since the income elasticity of M1 is likely to be less than one in the long run, the long-run expectation for its velocity is that it will increase.

Innovations in the financial sector and in the usage of money by non-financial economic agents in the economy have been very rapid in the last three decades, so that the money demand functions estimated with data including this period are often not stable. Further, it is even more rare to find the estimated functions for both narrow and wide definitions of money to be stable for a given country over a given period.

For the open economy, the existence of extensive CS could cause the monetary authority to lose control of the domestic money supply and increase the volatility of exchange rates under a floating exchange rate regime. It would increase the speculative pressures under fixed exchange rates. A common finding in estimations of the open-economy domestic money demand function is that the expected change in the exchange rate – which is the proxy on the return on holding foreign money relative to that on domestic money – is not significant in explaining domestic money demand. Such a finding has led to the conclusion that currency substitution tends to be extremely low, even in countries like Canada, in which the public often holds US dollars in currency or in US dollar bank deposit accounts. However, many studies also show that the return on foreign bonds is a significant positive determinant of domestic money demand. This could provide indirect evidence on CS: if foreign money balances provide monetary services as a medium of payments in the domestic economy, the decrease in their holdings due to an increase in the return on foreign bonds, has to be compensated by an increase in domestic money balances in order to maintain the desired holdings of all media of payments. This effect relies on the substitution between the domestic and foreign monies in their medium-of-payments role, while relying upon substitution between the foreign money and foreign bonds in portfolio allocation.

Several of the variables crucial to money demand estimation are not stationary. This is especially likely to be so for the monetary aggregates themselves, as well as for the income and wealth variables. It may or may not also be so for the interest rates in the particular data set. Consequently, the classical regression techniques do not yield unbiased and consistent coefficients. Cointegration analysis is an appropriate procedure in this case and has become quite common in recent years for estimating money demand functions. Its combination with error-correction modeling has the further advantage that the estimation yields both the long-run and the short-run demand functions.

Cointegration procedures represent an attempt to capture the long-run equilibrium relationship and there should exist a cointegrating vector if such a relationship is stable over the sample period. However, when long-run relationships are shifting due to innovations and the impact of the innovations has not been eliminated or somehow captured in the definition of the variables or the procedure used, the sample data would not incorporate a stable long-run relationship.

Money demand studies using cointegration techniques for data over the last few decades have provided a mixed bag of evidence about the existence of a cointegrating vector between money, income, interest rates and prices. The finding of such a vector has often been culled from the data by using different definitions of money, different interest rates and different periods. The last few decades have seen a mixed bag of very significant innovations related to money demand, so that the long-run money demand function must have been shifting. Consequently, cointegration studies, like earlier studies using the standard regression techniques, have not provided convincing evidence of the existence of a stable long-run money-demand function for the last few decades for Britain, Canada and the USA.

In studies where acceptable cointegration vectors have been established, an error-correction model has usually also been estimated. As expected, these studies show for quarterly data that the impact elasticities are relatively quite small and much smaller than the long-run elasticities, indicating that adjustments of money demand to changes in the independent variables take at least several quarters.

We have not differentiated between the demand functions estimated for the different segments of the economy, such as households, firms and financial institutions. There are numerous studies on these and the interested reader is encouraged to explore them. There is a significant difference between the demand for money by households and that by firms, especially large ones. In general, the former tends to be relatively more predictable than the latter.

Summary of critical conclusions

- The income elasticity of the demand for M1 is less than one.
- The income elasticity of M2 demand is higher than that of M1 demand and is sometimes estimated to be greater than one.
- The negative interest elasticity of the demand for money, no matter how it is defined, is now beyond dispute.
- Empirical studies do not show convincing evidence of the liquidity trap, even for data covering the 1930s.
- M1 has performed better than broader monetary aggregates during some periods and worse in others. Several recent studies have supported the use of M1 over M2 and broader aggregates.
- For the 1960s and 1970s, estimates based on a partial adjustment model often indicated a low impact (income) elasticity for the first quarter but a long-run elasticity close to one, indicating a slow adjustment of money demand to its long-run level.
- Financial innovations during the last three decades have rendered the money demand function unstable for this period. Numerous attempts and innovative variations in estimation have not established a stable demand function, with a specific form and invariant coefficients, for out-of-sample data.

❖ Most of the variables relevant to the money demand function have proved to be non-stationary. Therefore, most empirical studies now use cointegration analysis with an error-correction model. The latter is also used to judge causality between money and output.

Appendix

The ARDL model and its cointegration and ECM forms

As explained in Chapter 8, the regressors in an autoregressive distributed lag (ARDL) model include the lagged values of the dependent variable and the current and lagged values of the explanatory variables. Its estimating equation with p lagged values of the dependent variable and $q_j, j = 1, 2, \ldots, n$, values of the n explanatory variables is designated as an ARDL$(p, q_1, \ldots, q_n)$ and has the form:

$$\beta(L, p)y_t = \beta_0 x_0 + \sum_{j=1}^{n} \beta_j(L, q)x_{jt} + \mu_t \tag{32}$$

where L is the lag operator such that $L^i y_i = y_{t-i}$, x_0 is a constant and (L, p) and (L, q) are the lag polynomials:

$$\alpha(L, p) = 1 - \alpha_1 L^1 - \alpha_2 L^2 - \alpha_p L^p \tag{33}$$

$$\beta(L, q) = 1 - \beta_1 L^1 - \beta_2 L^2 - \alpha_q L^q \tag{33'}$$

In the long run, $y_t = y_{t-1} = \cdots = y_{t-p}$ and $x_{jt} = x_{jt-1} = \cdots = x_{jt-q}$, so that $L = 1$, $\alpha(1, p) = (1 - \alpha_1 - \alpha_2 - \alpha_p)$ and $\beta(1, q) = (1 - \beta_1 - \beta_2 - \beta_q)$ and the long-run relationship becomes:

$$y_t = \beta_0' + \sum_{j=1}^{n} \beta_j' x_{jt}' + v_t \tag{34}$$

where $\alpha_0' = \alpha_0/(\alpha(1, p), \beta_j = \beta_j'(1, q)/(\alpha(1, p), v_j = \mu_j/(\alpha(1, p)$. The error-correction equation of this ARDL model is:

$$\Delta y_t = \Delta\beta_0' - \sum_{i=2}^{p} \alpha_i' \Delta y_{t-i} + \sum_{j=1}^{n} \beta_{j0}' \Delta x_{it} - \sum_{j=1}^{n} \sum_{i=2}^{q} \beta_{i,t-j}' \Delta x_{j,t-i} - \alpha(1, p)\mathrm{ECM}_{t-1} + \eta_t \tag{35}$$

where:

$$\mathrm{ECM}_{t-1} = y_{t-1} - \hat{\beta} - \sum_{j=1}^{n} \beta_j x_{jt-1} \tag{36}$$

$\alpha(1, p)$ measures the speed of adjustment.

An illustration: a simple ARDL model

The simplest case of an ARDL model has only one explanatory variable x_1 and one-period lags, so that it is ARDL(1, 1). The estimation equation for this case is:

$$y_t = \alpha_0 + \alpha_1 y_{t-1} + \beta_0 x_{1t} + \beta_1 x_{1t-1} + \mu_t \tag{37}$$

where μ is white noise. The long-run relation between y and x_1 for this equation is obtained by setting $y_t = y_{t-1}$ and $x_t = x_{t-1}$, so that the long-run equation is:

$$y = \alpha_0/(1-\alpha_1) + \{(\beta_0+\beta_1)/(1-\alpha_1)\}x_{1t} + \{1/(1-\alpha_1)\}\mu_t \tag{38}$$

where $(\beta_0 + \beta_1)/(1 - \alpha_1)$ provides the long-run relationship between y and x. Further, in (38), replacing y_t by $(y_{t-1} + \Delta y_t)$ and x_{1t} by $(x_{1t-1} + \Delta x_{1t})$, we get:

$$\Delta y_t = \alpha_0 - (1-\alpha_1)y_{t-1} + \beta_0 \Delta x_{1t} + (\beta_0+\beta_1)x_{1t-1} + v_t \tag{39}$$

which is the short-run ECM representation of (37).

Further, from (37), we have:

$$\begin{aligned} y_t - \alpha_1 y_{t-1} &= \alpha_0 + \beta_0 x_{1t} + \beta_1 x_{1t-1} + \mu_t \\ (1-\alpha_1 L)y_t &= \alpha_0 + \beta_0 x_{1t} + \beta_1 x_{1t-1} + \mu_t \end{aligned} \tag{40}$$

Expanding $\{1/(1-\alpha_1 L)\}$, we have:

$$1/(1-\alpha_1 L) = (1 + \alpha_1 + {\alpha_1}^2 + \cdots)$$

Hence, (40) yields:

$$y_t = (1 + \alpha_1 + {\alpha_1}^2 + \cdots)\alpha_0 + (1 + \alpha_1 + {\alpha_1}^2 + \cdots)(\beta_0 x_{1t} + \beta_1 x_{1t-1} + \mu_t) \tag{41}$$

which is another way of stating (37). In this context, $(1 + \alpha_1 + \alpha_1^2 + \cdots)(\beta_0 + \beta_1) = (\beta_0 + \beta_1)/(1 - \alpha_1))$ provides the long-run relationship between y and x.

Review and discussion questions

1. Empirical studies of the demand for money in the last two or three decades have raised serious doubts about the stability of the money demand function. What were the main causes of this instability?

   Was the finding of instability related to the particular monetary aggregate used or did it occur across all aggregates? What were the main modifications made in the estimating equations and in the definitions of the variables in order to reach a stable money-demand function?
2. Specify the relevant relationships and discuss how the error-correction model can be used to assess causality between money and (i) nominal income, (ii) real output?
3. Specify the appropriate relationship and discuss the use of cointegration and error-correction estimates to judge (i) long-run, (ii) short-run neutrality of money.

4. Specify the appropriate relationship and discuss whether the cointegration and error-correction estimates can shed any light on the question of whether the deviations of output from its full-employment level are transitory and self-correcting, as the modern classical model asserts. If they were not such as to imply transitory and self-correcting deviations of output from its full-employment value, do they indicate any role for monetary policy to moderate these deviations?
5. How would you formulate your money demand function for estimation in an empirical study? Comment on the a priori relationships that you expect between your independent variables and money. Compare your demand function with some roughly similar and some different ones estimated in the literature.
6. For a selected country and using quarterly data, specify and estimate the money demand function. Check and correct for shifts in this function during the period of your study. Try the following variations of the independent variables:

   (i) Expected income and permanent income for the scale variable.
   (ii) Two different interest rates, one short-term and the other medium-term.
   (iii) A proxy for the expected change in the exchange rate.
   (iv) Also, do your estimations using the following techniques:
      (a) least squares estimation, with a first-order PAM;
      (b) cointegration with an error-correction model.
   (v) Discuss your choice of the functional form of the money demand function and your choice of the variables and the econometric techniques used, as well as the data and econometric problems you encountered.
   (vi) Discuss your results, their plausibility and consistency with the theory, and their robustness.

7. Discuss: shifts in the estimated coefficients of the money demand function are as likely to be due to shifts in monetary policy (supply side shifts) as about money demand behavior. (This question relates to the identification of demand versus supply functions.)
8. What are the reasons for requiring the use of cointegration techniques in money demand estimation? What would be the disadvantages of using ordinary least squares for such estimation? If you use both and obtain different estimates, which would you rely on, and why?
9. Are there any conceptual problems with the application of the cointegration techniques to money-demand functions, or can the estimates from such techniques be relied upon? In particular, how can you make certain that the estimated cointegration vector is the money demand and not the money supply function or a reduced-form relationship between money demand and money supply?
10. Specify the Taylor rule. Discuss its estimation by cointegration and error-correction techniques, specifying and justifying your choices of the dependent and explanatory variables.
11. In the context of interest-rate targeting (as against monetary targeting), suppose you wanted to estimate a St Louis equation for (i) nominal income, (ii) real output, but with the interest rate as the monetary policy variable. Specify the appropriate equation. Discuss its estimation by cointegration and error-correction techniques.
12. Conduct an empirical study along the lines suggested in the preceding question for a country of your choice, and discuss your findings for the effectiveness of monetary policy pursued through interest rates.

13. Design a study to judge whether a central bank of a country uses the interest rate or the money supply as its operating target of monetary policy. What relationships and tests can you use for this purpose?

References

Baba, Y., Hendry, D.F., and Starr, R.M. "The demand for M1 in the U.S.A., 1960–1988." *Review of Economic Studies*, 59, 1992, pp. 25–61.

Bae, Y., Kakkar, V., and Ogaki, M. "Money demand in Japan and nonlinear cointegration." *Journal of Money, Credit and Banking*, 38, 2006, pp. 1659–67.

Coenen, G., and Vega, J.L. "The demand for M3 in the Euro area." *Journal of Applied Econometrics*, 16, 2001, pp. 727–48.

Cuthbertson, K. "Modelling the demand for money." In C.J. Green and D.T. Llewellyn, eds, *Surveys in Monetary Economics*, Vol. 2. Cambridge, MA: Blackwell, 1991.

Davidson, R., and Mackinnon, J.G. *Estimation and Inference in Econometrics*. New York: Oxford University Press, 1993.

Engle, R.F., and Granger, C.W.J. "Cointegration and error-correction: representation, estimation and testing." *Econometrica*, 55, 1987, pp. 251–76.

Fair, R.C. "International evidence on the demand for money." *Review of Economics and Statistics*, 69, 1987, pp. 473–90.

Feige, E.L. "Expectations and adjustments in the monetary sector." *American Economic Review*, 57, 1967, pp. 462–73.

Feige, E.L., and Pearce, D.K. "The substitutability of money and near-monies: a survey of the time-series evidence." *Journal of Economic Literature*, 15, 1977, pp. 439–70.

Goldfeld, S.M. "The demand for money revisited." *Brookings Papers on Economic Activity*, 3, l973, pp. 576–638.

Goldfeld, S.M., and Sichel, D.E. "The demand for money." In B.M. Friedman and F.H. Hahn, eds, *Handbook of Monetary Economics*. Amsterdam: North-Holland, 1990, Volume 1, Chapter 8, pp. 299–356.

Hafer, R.W., and Jansen, D.W. "The demand for money in the United States: evidence from cointegration tests." *Journal of Money, Credit and Banking*, 23, 1991, pp. 155–68.

Haug, A.A. "Canadian money demand functions: cointegration-rank stability." *The Manchester School*, 74, March 2006, pp. 214–30.

Johansen, S. "Statistical analysis of cointegration vector." *Journal of Economic Dynamics and Control*, 12, 1988, pp. 231–54.

Johansen, S. "Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models." *Econometrica*, 59, 1991, pp. 1551–80.

Johansen, S., and Juselius, K. "Maximum likelihood estimation and inference on cointegration with application to the demand for money." *Oxford Bulletin of Economics and Statistics*, 52, 1990, pp. 169–210.

Judd, J.P., and Scadding, J.L. "The search for stable money demand function: a survey of the post-1973 literature." *Journal of Economic Literature*, 20, 1982, pp. 993–1023.

Keynes, J.M. *The General Theory of Employment, Interest and Money*. London and New York: Macmillan, 1936.

Miller, S.M. "Money dynamics: an application of cointegration and error-correction modeling." *Journal of Money, Credit and Banking*, 23, 1991, pp. 139–54.

Miyao, R. "Does a cointegrating M2 demand relation really exist in the United States?" *Journal of Money, Credit and Banking*, 28, 1996, pp. 365–80.

Mulligan, C., and Sala-i-Martin, X. "Extensive margins and the demand for money at low interest rates." *Journal of Political Economy*, 108, 2000, pp. 961–91.

Pesaran, M.H., Shin, Y., and Smith, R.J. "Bounds testing approaches to the analysis of level relationships." *Journal of Applied Econometrics*, 16, 2001, 289–326.

Poloz, S.S. "Simultaneity and the demand for money in Canada." *Canadian Journal of Economics*, 13, 1980, pp. 407–20.

Sriram, S.S. "Survey of literature on demand for money: theoretical and empirical work with special reference to error-correction models." *IMF Working Paper* no. 64, 1999.

Sriram, S.S. "A survey of recent empirical money demand studies." *IMF Staff Papers*, 47, 2000, 334–65.

Part IV

Monetary policy and central banking

10 Money supply, interest rates and the operating targets of monetary policy

Money supply and interest rates

This is the first of three interrelated chapters on monetary policy and central banking. It starts by examining the goals and operating targets of monetary policy. The two major operating targets of monetary policy are the money supply and the interest rate.

This chapter then focuses on the determination of the money supply. While macroeconomic models tend to simplify by assuming that the money supply is exogenously determined, the private sector in the form of the banks, households and firms also influences the money supply.

Key concepts introduced in this chapter

- Targeting inflation or its deviation from a desired inflation rate
- Targeting output and unemployment
- Interest rate as an operating target
- Monetary base
- Currency ratio
- Demand deposit ratio
- Free reserves
- Excess reserves
- Required reserves
- Discount/bank rate
- Mechanical theories of the money supply
- Behavioral theories of the money supply

This is one of three chapters on some of the central issues of monetary policy. It starts with the relationships among the goals, intermediate targets and operating targets of monetary policy and examines the theoretical justification as well as the implications of adopting different targets. It then considers the issue of whether the central bank should use the money supply or the interest rate as its major monetary policy instrument. It then narrows its focus to the determination of the money supply in the economy, so as to complement the extensive treatment of money demand in the preceding chapters.

Sections 10.1 and 10.2 present the links between the goals and targets of monetary policy. Sections 10.3 to 10.5 examine the main operating targets of monetary policy commonly

used by central banks and their justification from macroeconomic analysis.[1] Sections 10.6 to 10.8 present the determination of the money supply. Section 10.9 covers the application of cointegration analysis and error-correction modeling to money supply. Section 10.10 considers the central bank's choice between the monetary base and the interest rate as alternative operating targets.

Stylized facts on the goals and operating targets of monetary policy

The stylized facts on monetary policy depend on the behavior of the central bank and the structure of the economy. Among these facts are:

1 The central bank has more than one goal. Among its goal variables are output and its growth rate, unemployment, inflation, etc. Currently many central banks focus on reducing the deviation of output from its full-employment level and of inflation from a target level, with a trade-off between them, as in a Taylor rule.
2 The target inflation rate for many central banks now is a low inflation rate, often in a range of 1 percent to 3 percent.
3 The operating target of monetary policy can be a monetary aggregate or an interest rate. A monetary aggregate was selected for this purpose in some past periods and is still in use by some central banks. Currently, many central banks in the developed economies focus on an interest rate as their primary operating target.
4 The central bank does not control the money supply directly but has to use its instruments, such as the monetary base, for indirectly controlling the money supply.

10.1 Goals, targets and instruments of monetary policy

The eventual purpose of monetary policy is to achieve certain national goals. These have historically included full employment (or a low unemployment rate), full-employment output (or a high output growth rate), a stable price level (or a low inflation rate), a stable exchange rate (or a desirable balance of payments position), etc. These variables are simply referred to as "*goals*" or as "ultimate goals" of monetary policy. However, the central bank cannot achieve these goals directly by its monetary policy *instruments*, which are variables that it can operate on directly. Among the instruments available to the central bank are open-market operations and changes in its discount/bank rate at which it lends to commercial banks and other bodies. These determine the economy's monetary base. In many countries, the central bank can also change the required reserves (i.e. the minimum reserves the commercial banks must hold against the public's deposits with them), which changes the "monetary base multiplier" (i.e. the money supply per dollar of the monetary base). These measures serve to change the money supply in the economy. Another monetary policy instrument is the overnight loan rate (called the federal funds rate in the USA) in the market for reserves, whose operation induces change in various interest rates in the economy. The next chapter provides further information on the goals and instruments of monetary policy.

1 This section requires some prior knowledge of the IS–LM and IS–IRT models of aggregate demand from macroeconomic courses. A review of these models is provided in Chapter 13.

Besides the concepts of goals and instruments, other concepts relevant to monetary policy are those of targets, operating targets and guides. We can broadly define a *target* variable as one whose value the policy maker wants to change.[2] An *operating target* variable is one on which the central bank can directly or almost directly operate through the instruments at its disposal. A *guide* is a variable that provides information on the current and future state of the economy.

Between the goals and instruments of monetary policy lie layers of intervening variables. For example, suppose the central bank wants to reduce the inflation rate. To do so, it needs to reduce aggregate demand in the economy. The reduction in aggregate demand usually requires a reduction in investment and/or consumption, which requires an increase in market interest rates. Depending on the analysis, discussion or author, these intervening variables can be referred to as intermediate targets, operating targets or even as instruments. Since a target variable is one whose value the central bank seeks to influence or control by the use of the tools at its disposal, any of the intervening variables between the goals and instruments can be referred to as a target variable. In the preceding example, aggregate demand is an intermediate variable or target, which the central bank wants to alter by using the intermediate targets of the money supply and/or interest rates which, in turn, can be altered by changes in the monetary base and the discount rate. Note that the word "target" can also be used to indicate a desirable value of a goal (e.g. inflation) or of an intermediate variable (e.g. the money supply and market interest rates).

Given the preceding discussion, Table 10.1 provides a rough classification of monetary policy instruments, operating targets, intermediate targets and goals.

While Table 10.1 provides some guidance on the roles and sequence of the various monetary policy variables, there is no hard and fast rule for its classification. The central bank uses its tools to hit its operating targets, with the intention of manipulating the intermediate targets, which are the final ones of the financial system, in order to achieve its goals. Note that lags enter at each stage of this process, and both the individual lag and the overall lag tend to vary. Further, the duration of the lags and the final impact are not usually totally predictable.

Table 10.1 Monetary policy tools, target and goals

Policy instruments	*Operating targets*	*Intermediate targets*	*Goals*
Open-market operations	Short-term interest rates	Monetary aggregates (M1, M2, etc.)	Low unemployment rate
Discount rate	Reserve aggregates (monetary base, reserve, nonborrowed reserves, etc.)	Interest rates (short and long term)	Low inflation rate
Reserve requirements			Financial market stability
			Exchange rates
		Aggregate demand	

2 Under this broad definition, targets can be ultimate ones (final goals, such as output and unemployment), intermediate ones (such as the money supply or the interest rate) or operating ones (such as the monetary base or the discount rate). Since a given variable can fall in any one of these categories, there is no hard and clear-cut separation between these categories.

10.2 Relationship between goals, targets and instruments, and difficulties in the pursuit of monetary policy

Several issues arise in the selection and use of goals, intermediate variables and operating targets or instruments by the monetary authorities. Among these are:

1 Are the relationships between the ultimate goal variables, intermediate variables and operating targets stable and predictable?
2 Can the central bank achieve the desired levels of the operating targets through the instruments at its disposal?
3 What are the lags in these relationships, and, if they are long, can the future course of the economy be reasonably well predicted?

To illustrate these points, let the relevant relationships be:

$$y = f(x; \Psi) \tag{1}$$

$$x = g(z; \theta) \tag{2}$$

where:

y = (ultimate) goal variable
x = intermediate target
z = policy instrument or operating target
Ψ, θ = sets of exogenous variables

The above equations imply that:

$$y = h(z; \Psi, \theta) \tag{3}$$

so that z can be used to achieve a desired value of y. However, this can be done reliably only if the functional forms f and g are known and these are stable univalued functions.[3] In practice, given the complex structure of the real-world economies, as well as the existence of uncertainty and lags in the actual relationships, the precise forms of f, g and h are often only imperfectly known at the time the decisions are made. Further, the coefficients in these relationships may be subject to stochastic changes. In addition, given the lags in the economy, the policy maker also needs to predict the future values of the coefficients and the exogenous variables – again, usually an imprecise art.

Hence, the precision and clarity implied by (3) for the formulation of monetary policy and its effects is misleading. In many, if not most instances, the impact of a change in most of the potential operating variables on the ultimate goals is likely to be imprecise, difficult to predict and/or unstable. This makes the formulation of monetary policy an art rather than a science and cautions against attempts to use monetary policy as a precise control mechanism for "fine-tuning" the goals of such policy.

Another common problem with most target variables is that they are endogenous and their values depend on both demand and supply factors, so that the exogenous shocks to them could come from either demand or supply shifts. The policy maker may want to offset the effect of changes in some of these factors but not in all cases, so that it needs to know the source of such changes before formulating its policy.

3 See also Chapters 9, 14–17 for material relevant to this discussion.

10.3 Targets of monetary policy

The two main *operating targets* usually suggested for monetary policy are:

- monetary aggregates;
- interest rates.

The two main *targets* of monetary policy highlighted in the recent literature are:

- inflation rate (or the price level),[4] or its deviation from a desired value;
- output, or its deviation from the full-employment level.

There are also other variables that are sometimes used or proposed as the intermediate targets of monetary policy. Among these is aggregate demand (or nominal national income) and, in the case of relatively open economies, the exchange rate or the balance of payments. For the sake of brevity, this chapter discusses only the relative merits and demerits of monetary aggregates and the interest rate as the chief operating target or instruments. It also presents some discussion of the price level and the inflation rate, and the output gap, as the targets of monetary policy.

10.4 Monetary aggregates versus interest rates as operating targets

This section relies upon students' prior knowledge of the IS–LM macroeconomic model (otherwise, see Chapter 13) to distinguish between the relative merits of using the money supply versus interest rates as the operating target of monetary policy. The choice between monetary aggregates and the interest rate depends critically upon the policy objective of the central bank and the structure of the economy. The following analysis, adapted[5] from that in Poole (1970), takes this objective to be control of aggregate demand,[6] since the central bank can only influence output and inflation, which are its final goal variables, through manipulation of aggregate demand. It further assumes that the structure of the economy can be represented by the IS–LM analysis and diagram. This diagram has aggregate real demand y on its horizontal axis and the real interest rate r on its vertical axis. The commodity market equilibrium is shown by the IS curve and the money market equilibrium is shown by the LM curve. Their intersection determines real aggregate demand at the existing price level.

4 Price stability is also sometimes designated as a primary policy goal. However, even in such a context, the justification given for it is that price stability promotes the achievement of full employment and output growth.

5 This adaptation takes the control of aggregate demand rather than that of real output as the objective of monetary policy. Poole had treated the two as identical under the assumption that the price level was constant. Since this assumption is both unnecessary and unrealistic, our discussion is based on the objective of minimizing the variance of aggregate demand rather than of output.

6 However, note that the literature does also include other goal variables. One of these is the variance of the money supply, with the choice between the monetary base and the interest rate depending on which instrument minimizes this variance under shocks to money demand and money supply. In this analysis, when the interest rate is the policy instrument and the money supply is accommodated to money demand, shocks to both money demand and money supply affect the money supply. However, when the monetary base is the policy instrument, only the shock to the monetary base (to money supply) multiplier determines fluctuations in the money supply. Therefore, controlling the monetary base leads to smaller fluctuations in the money supply. We do not regard the objective of minimizing the variance of money supply to be an appropriate goal, and do not present the analysis related to it.

Therefore, the choice between the monetary instruments hinges on the question: which instrument provides better control over aggregate demand in the IS–LM framework? Our analysis implicitly assumes the Fisher equation for perfect capital markets and an expected inflation rate of zero, so that the nominal interest rate R is identical with the real interest rate r.

Since the IS–LM analysis has not yet been mathematically covered in this book, this chapter presents only the diagrammatic analyses of monetary versus interest rate targeting. Its mathematical version is presented in Chapter 13, which could be read at this point.

10.4.1 Diagrammatic analysis of the choice of the operating target of monetary policy

Shocks arising from the commodity market

The IS equation and curve encompass the various components of expenditures, such as consumption, investment, exports, fiscal deficits, etc., in the economy (see Chapter 13). Several of these are volatile, with investment often being the most volatile component of expenditures. Shifts in any of these components shift the IS curve in the IS–LM diagram.

Our analysis starts with the initial equilibrium shown by point a, with coordinates (r_0, y_0), in Figure 10.1a. Assume that the central bank targets the money supply and holds it constant through open market operations or by the use of some other instruments. Shocks to the IS curve[7] would then change both r and y. To illustrate, if a positive shock shifts the IS curve from IS_0 to IS_1, aggregate demand will increase from y_0 to y_1 and the interest rate rise from r_0 to r_1. Similarly, a negative shock, occurring, say, in the following period, which shifts the IS curve to IS_2, will lower aggregate demand to y_2 and the interest rate to r_2.

Compare this result with the impact of the same shock if the interest rate had been targeted. This is shown in Figure 10.1b, where the interest rate is assumed to be held fixed by the authorities at the target rate $\underline{r}_0$, where the underline indicates that it is exogenously set by the central bank. The shifts in the IS curve, first to IS_1 and then to IS_2, will produce movements

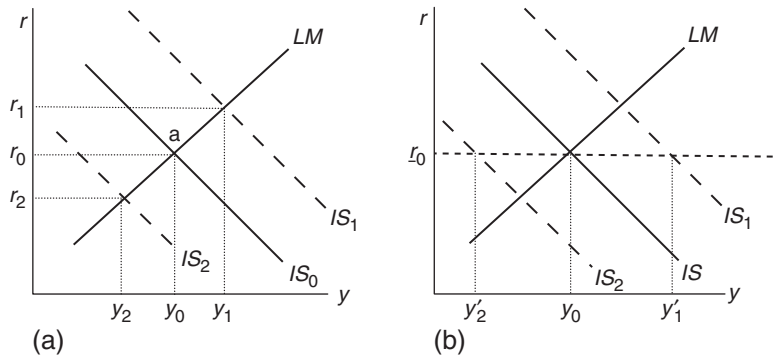


Figure 10.1

7 Shocks originating in the commodity market are to consumption, investment, exports and government expenditures. Of these, investment is considered to be the most volatile element.

in aggregate demand, first to y_1' and then to y_2'. This fluctuation between y_1' and y_2' is clearly greater than between y_1 and y_2 in Figure 10.1a, so that targeting the interest rate produces greater fluctuations in aggregate demand than money supply targeting if the exogenous shocks emanate from the commodity market. Note that such shocks do not produce changes in the interest rate, since that is being held constant through monetary policy.

Shocks arising from the money market

Now assume that the exogenous shocks arise only in the money market while there are no shocks in the commodity market, so that the IS curve does not shift. Such exogenous shocks in the money market can be to either money demand or money supply, and shift the LM curve.

Money supply targeting would stabilize the money supply,[8] so that disturbances to it do not have to be considered, but not the money demand. Now suppose that money demand decreases. Given the targeted money supply, the decrease in the money demand will shift the LM curve in Figure 10.2 to the right to LM_1 and increase aggregate demand from y_0 to y_1. Assume that the next period's shock to the money demand increases it and shifts the LM curve to LM_2, so that aggregate demand falls to y_2. The aggregate demand fluctuations are then from y_1 to y_2 and the interest rate fluctuations are from r_1 to r_2.

For interest rate targeting, assume that the real interest rate had been set at r_0, as shown in Figures 10.3 and 10.4. Figure 10.3 shows the initial demand curve for nominal balances as M^d and the initial supply curve as M^s, with the initial equilibrium interest rate as r_0 and the initial money stock as M_0. Now suppose that the money demand curve shifts from M_0^d to M_1^d. Since the interest rate is being maintained by the monetary authority at $\underline{r}_0$, the monetary authority will have to increase the money supplied from M_0 to M_1. The money stock therefore adjusts endogenously through an accommodative monetary policy to the changes in money demand.

In the IS–LM Figure 10.4, a reduction in the money demand would shift the LM curve to the right from LM_0 to LM_1. However, given that the monetary authority maintains the

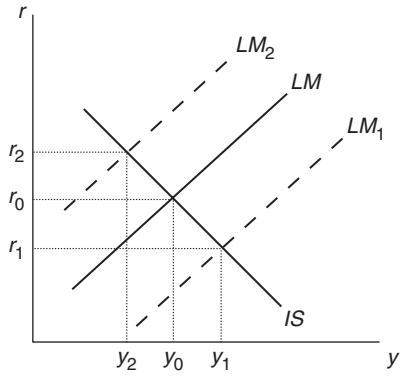


Figure 10.2

8 However, monetary base targeting usually will not do so.

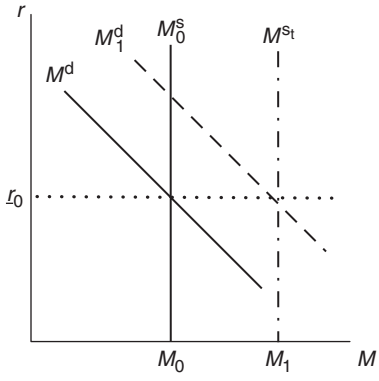


Figure 10.3

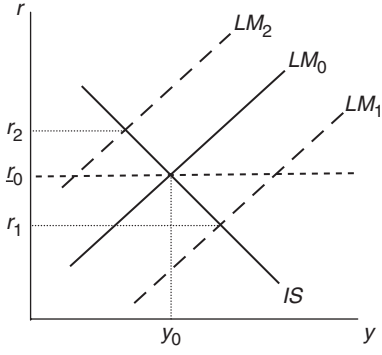


Figure 10.4

interest rate at $\underline{r}_0$, the aggregate demand y_0 in this figure will be determined by the intersection of the IS curve and a horizontal line at the target interest rate $\underline{r}_0$. This is so because the exogenous shift in the LM curve from LM_0 to LM_1 leads the central bank to undertake an accommodative money supply decrease sufficient to shift this curve back to LM_0. Hence, in spite of any exogenous changes in money demand, aggregate demand would remain at y_0 (and the interest rate at $\underline{r}_0$). Hence, comparing the implications from Figures 10.2 and 10.4, monetary targeting will allow greater fluctuations in aggregate demand and interest rates than interest-rate targeting when the exogenous shifts arise from money demand.

This conclusion poses a problem for the policy maker since both types of shocks occur in the real world. Therefore, the monetary authority has to determine the potential source of the dominant shocks to the economy before making the choice between monetary and interest rate targeting. This is not easy to determine for the future, nor need the same pattern of shocks necessarily occur over time. Further, since both types of shock do occur, each policy will reduce or eliminate the impact of some types of shocks but not of others.

While many central banks had, for a few years during the late 1970s and sometimes in the early 1980s, favored monetary targeting, the common practice currently is to set interest rates.

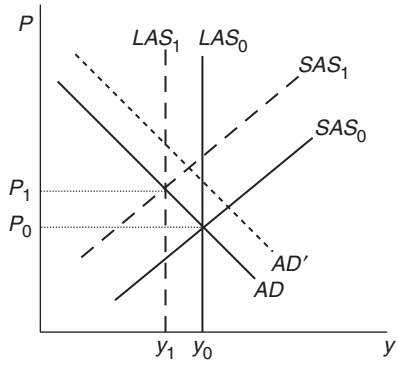


Figure 10.5

This implies, in the context of the preceding analysis, that the dominant sources of shocks are expected to be in the monetary sector.

10.4.2 Analysis of operating targets under a supply shock

For the analysis of operating targets under supply shocks, we start by changing the objective function from stabilization of aggregate demand y^d to stabilization of the price level P or/and real output y. Figure 10.5 shows the aggregate demand (AD) curve and the short-run (SAS) and long-run (LAS) aggregate supply curves for the economy. The initial equilibrium is at (y_0, P_0). An anticipated *negative permanent* supply shock will shift the supply curves from SAS_0 to SAS_1 and LAS_0 to LAS_1. First, consider the short-run effect of the fall in supply to SAS_1. Prices rise from P_0 to P_1, while output falls from y_0 to y_1. The rise in prices will decrease the real money supply and shift the LM curve to the left (for instance, from LM_0 to LM_2 in Figure 10.4), so that the interest rate will rise (from r_0 to r_2). Monetary targeting will leave the money supply unchanged and therefore leave the new equilibrium at y_1, P_1 and r_2.

But an interest rate target at $\underline{r}_0$ will cause the central bank to increase the money supply to prevent the interest rate from rising. This will increase aggregate demand and cause a policy-induced shift from AD to AD′ in Figure 10.5. The result will be a further increase in prices but the fall in output will be partly or wholly (depending on the induced demand increase) offset in the short run. The relevant intersection is that of SAS_1 and AD′. Hence, in the short run, interest-rate targeting is more inflationary than monetary targeting but compensates for this by limiting the fall in output.

Now consider the long-run analysis with the shift from LAS to LAS_1. In this case, interest-rate targeting will cause a continual increase in the money supply and the price level, without any beneficial offset in terms of output or the maintenance of the interest rate at $\underline{r}_0$. Therefore, for permanent supply shocks, monetary targeting is clearly preferable in the long run, whereas interest-rate targeting involves a cumulative inflationary process.

Monetary aggregates as targets in practice

Milton Friedman and the 1970s monetarists, belonging to the St Louis school, had argued that because of the existence of both a direct and an indirect transmission mechanism from

the money supply to aggregate expenditures, the money supply rather than interest rates provided better control over the economy. Partly as an outcome of this advice, most countries – including the USA, Britain and Canada – switched to the targeting of monetary aggregates after the mid-1970s (though only until the early 1980s). The monetary aggregates often suggested as targets were M1 or M2 – and M4 in Britain – though sometimes even broader targets were also considered.

Monetary aggregate targeting was predicated on the belief that the relationship between such a target and aggregate demand was stable and had a short and predictable lag. This was certainly the finding of the studies done by the St Louis school. Monetary targets were pursued in the late 1970s and early 1980s by the monetary authorities in the USA, Canada and UK. However, the functional relationships between the monetary variables and aggregate expenditures, let alone the rate of inflation, proved to be unstable, so that they had been abandoned by the 1990s in each of these countries. Among the reasons for this instability were financial innovations and changes in the payments technology occurring in recent decades.[9] In terms of experience during the late 1970s and 1980s, direct targeting of monetary aggregates increased both the level and the volatility of interest rates considerably, with the latter considered by many economists to be destabilizing for the economy. Attempts to control the monetary or reserve aggregates directly, as a way of controlling the economy, were abandoned by most central banks in the early 1980s in favor of interest-rate targets as the control variable. This is not to say that the monetary aggregates are not monitored and the changes in them not considered in formulating monetary policy. However, for most central banks, they have ceased to be the main operating targets.

Interest rates as targets in practice

Monetary policy acts through interest rates on spending, so that the interest rates are closer in the chain of influence on spending. Hence, they are more reliable and more appropriate indicators of the need for action than are the various measures of money supply and the monetary base. In line with this, in financially developed economies such as those of the USA, Canada and the UK, the central banks believe that interest rates are a major indicator of the performance of the economy and tend to use them as the preferred guide and operating target of monetary policy.[10]

There are several measures of interest rates that may be considered, with the usual selection for operating purposes being of short-term nominal, rather than long-term or real, rates of interest. Historically, the measure commonly used for this purpose used to be the Treasury bill rate. As discussed later in Chapter 11, more recently the USA, UK and Canada have used an overnight loan rate as an operating target. These countries have well-developed markets for overnight loans among financial institutions, with this market serving as the market for the excess reserves of banks. This market for reserves is known as the Federal Funds market in the United States and the overnight loan market in Canada and the UK. Such a rate reflects the commercial banks' demand and supply conditions for reserves. The central bank's policy actions on the monetary base immediately affect the commercial banks' demand and supply

9 These financial innovations included the payment of interest on checking accounts and the increasing degree of substitution between M1 and near-monies, telephone and on-line banking, etc.

10 In this context, see the discussion in Chapter 11 on the Monetary Conditions Index used by the Bank of Canada.

of reserves, thereby changing the overnight interest rate and starting a chain of reactions on other interest rates, and through these on the borrowing and lending, investment and consumer spending, etc., in the economy. A higher rate means that banks are relatively loaned up and a lower rate means that banks have relatively large free reserves, so that they can increase loans of their own volition.

Problems with the use of interest rates in managing the economy

The observed interest rates are equilibrium rates, so that changes in them could reflect either changes in demand or in supply conditions or both. Therefore, a rise in the interest rates may be due to an increase in the demand for loanable funds or a decrease in their supply, but the central bank may wish to take offsetting action in only one of these cases. For example, interest rates rise during an upturn in the business cycle. The central bank may not wish the upturn to be curbed by a decreased supply of funds but also may not wish to offset the stabilization effect of interest rates due to an increase in their demand. But changes in the equilibrium interest rates do not by themselves provide adequate information as to the causes of their rise and therefore as to the policy actions that should be undertaken. Consequently, central banks in practice supplement information on interest rates with other information on demand and supply conditions before making their policy decisions.

A problem with using interest rates as an operational target is that the central bank can determine the general level of interest rates but not equally well control the differentials among them. Examples of these differentials are the loan-deposit spread of commercial banks, and the spread between deposit rates and mortgage rates, if the latter are variable. Spreads depend upon market forces and can be quite insensitive or invariant to the central bank's discount rate. Financial intermediation in the economy is more closely a function of such differentials than of the level of interest rates, so that the ability of the central bank to influence the degree of financial intermediation through its discount rate and the overnight loan rate for reserves becomes diluted.

Among other problems is the lag in the impact of changes in the interest rate on aggregate demand in the economy. Among the reasons for such lags are the costs of adjustment of economic variables such as the capital stock and planned consumption expenditures, and the indirect income effects of changes in interest rates. There are two aspects of this lag: its length and variability. The former is often assessed at about six quarters to two years in the United States, Britain and Canada. While there is agreement that there is some variability in the length of the lag, there is no consensus on whether it is so long that changes in interest rates, intended to be stabilizing, can prove to be destabilizing. Within the lag, the *impact* effect (within the same quarter) of interest rate changes on real aggregate demand is estimated to be quite low, while the *long-run* effect is now believed to be very significant.

The actual use of interest rates for stabilization has often been found to be "too little, too late" – though this is usually a result of uncertainty about the need for and the lags in the effects of monetary policy. This results in its cautious use, no matter what operational or indicator variable is used. Given the duration of lags and the uncertainty at any time about the position of the economy in the business cycle, past experience does indicate that central banks often change the interest rates later and in smaller steps than really needed. An initial change is, therefore, often followed by many more in the same direction over several quarters.

Money supply under an interest rate target

In market economies, the use by the central bank of the interest rate as its major instrument of monetary policy does not imply that it can ignore the money supply altogether. Interest rates are determined in financial markets, so that if the central bank were to lower its interest rate and not provide the supporting required increase in the money supply, it would find that the market rates will diverge from its desired ones, so that the intended effects on expenditures will not be achieved. Hence, an interest-rate policy must be accompanied by an appropriate money supply. This topic is addressed in the macroeconomic context in Chapter 13.

10.5 The price level and inflation rate as targets

Targeting the price level

Current discussions of monetary policy often refer to inflation or price targeting as the goal of monetary policy. A stable price level or a low inflation rate is sometimes proposed as the *ultimate goal* of monetary policy. For this, it is argued that money is neutral in the long run, so that the central bank cannot change the level and path of full-employment output, nor should it attempt to do so since such an attempt will only produce inflation. Under this neutrality argument, what the central bank can do is to ensure a stable value of money, so that its target should be in terms of the price level or the rate of inflation. Further, a fairly stable price level reduces the risks in entering into long-term financial contracts and fixed real investments, and promotes the formulation and realization of optimal saving and investment, which in turn increase output and employment. By comparison, high and variable inflation rates inhibit economic growth by introducing uncertainty into long-term financial contracts and investment.

For the following analyses of the price level and the inflation rate as the monetary authorities' target, we leave aside the comparison of monetary versus interest rates as targets and focus on aggregate demand as the variable in the control of the monetary authority, and assume that it will adopt the appropriate instrument to achieve the desired level of aggregate demand. Further, since our analysis is short run, we use a positively sloping short-run aggregate supply curve rather than a vertical long-run one.

Figure 10.6 assumes that there is a positive *demand shock* such that the AD curve shifts to AD_1. If the monetary authorities stabilized prices at P_0, output would remain unchanged at y_0. To achieve this under monetary targeting, the monetary authority would pursue a compensatory decrease in the money supply or an increase in the interest rate to shift aggregate demand back to AD. Under interest-rate targeting, they would raise the interest rate to achieve the same effect. The net effect of such a monetary policy would stabilize both the price level and output in the event of exogenous shocks from the money or commodity markets.

Figure 10.7a shows the effects of a negative supply shock such that the short-run aggregate supply curve SAS shifts from SAS_0 to SAS_1. This will produce an increase in the price level from P_0 to P_1 and a decrease in output from y_0 to y_1. Since the price level is not an operational variable under the direct control of the central bank, the bank would have to achieve price stability through a reduction in aggregate demand, which requires a contraction of the money supply or a rise in interest rates such that AD is made to shift to AD′. This will, however, decrease output from y_0 at P_0 to y_1 at P_1 due to the supply shock and then to y_1' due to the contractionary monetary policy and its implied shift of the AD curve to AD′. Hence, the

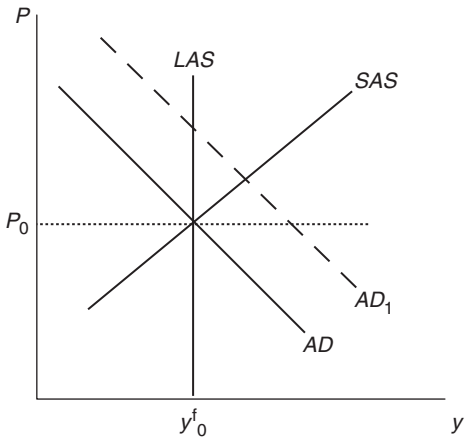


Figure 10.6

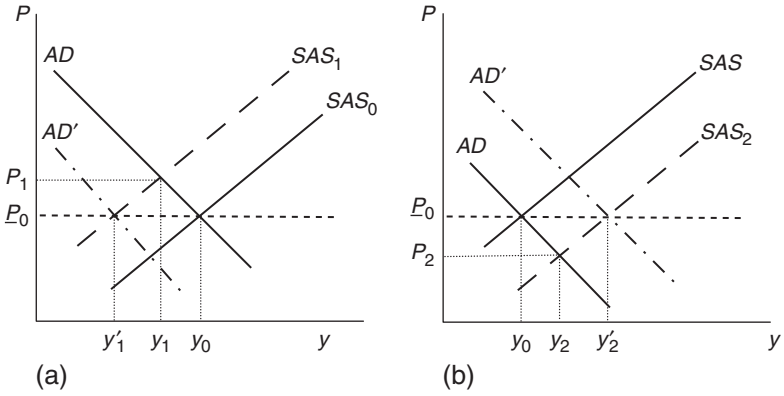


Figure 10.7

contractionary monetary policy would have increased the fall in output over that which would have occurred if the monetary policy had not been pursued.

Similarly, suppose that the aggregate supply shock had been a positive one, as shown in Figure 10.7b. This would shift the SAS curve to the right from SAS_0 to SAS_2, resulting in the increase in output from y_0 to y_2 and the decrease in prices from P_0 to P_2. The central bank could increase aggregate demand to stabilize the price level at P_0, but this would mean an expansionary monetary policy which shifts the AD curve to AD′ and further increases output to y'_2. Price stabilization has, therefore, again increased the fluctuation in output.

Therefore, given the aggregate supply curve as being positively sloped and short run, the pursuit of price stability in the face of supply-side fluctuations has the cost of increasing the instability of output – and, therefore, of unemployment – in the economy. We leave it to the reader to adapt the analysis to the case of a vertical long-run supply curve.

Targeting the inflation rate

A low inflation rate, say in the 1 percent to 3 percent range, is generally considered to be effectively consistent with price-level stability, with the increase in prices merely reflecting the continual improvements in existing products and the introduction of new ones. Further, a positive but low rate of inflation is often considered to be beneficial for the economy, particularly in the labor market where it gives firms the flexibility to respond to shifts in the relative demand or supply of different products and types of workers, as well as shifts over time in the performance of a given worker. On the latter, firms can respond to small declines in productivity without having to reduce nominal wages, which creates industrial unrest, of workers whose real wage would fall. Inflation, as well as labor productivity increases, overcomes the societal norm of downward nominal wage rigidity. As against this beneficial so-called "grease effect" of inflation, errors in inflationary expectations can lead to a nominal wage being set in explicit and implicit labor contracts that result in a real wage higher or lower than the one that ensures full employment in the economy. This so-called "sand effect" occurs because of the two stages of wage negotiation and employment/production relevant to the derivation of the expectations-augmented Phillips curve (see Chapter 14). Such errors in inflationary expectations are less likely to occur with low, pre-announced and credible inflation targets than with high ones. Therefore, many central banks and economists generally believe that a low, pre-announced and credible inflation target improves the real performance of the economy in both the short and the long run.

Note that the inflation rate is not an operating target, since the monetary authority cannot directly change it. To maintain a target range for the inflation rate, the central bank will have to operate on the monetary aggregates and/or interest rates. Its success or failure will depend on the predictability of the relationships between the rate of inflation and these variables. Since the central banks of many countries have pursued a low inflation rate as a goal for more than a decade, a considerable amount of evidence has accumulated on it. This evidence shows that this goal has, in general, resulted in a reduction in the actual inflation rates. However, given the aggressive pursuit of this goal, this is not a surprising finding. However, as shown in the analysis above of price level targeting, targeting the price level alone tends to cause increased fluctuations in output and unemployment. This does not seem to have occurred in the past two decades, perhaps because central bank policies have followed not the single goal of price stability or a low inflation but a Taylor rule, which addresses both the output gap and the deviation of inflation from its target level. Chapters 11 and 15 also address this point.

A low inflation target versus a stable price-level target

Under the price-level target, if the actual price level falls below or rises above the target level, future policies would have to aim to bring it back to the target level. Hence, rising prices would have to be offset by future deflationary policies to make the price level return to its target level. Such a deflationary policy usually imposes costs in output and unemployment. By comparison, targeting the inflation rate allows the central bank to ignore one-time shifts in the price level, such as those due to changes in indirect tax rates, a shift in relative prices or an adjustment in the exchange rate, etc.

In addition, many economists believe that the public more easily relates to a low inflation rate target that remains constant over time and to the policies needed to maintain it, than to a price-level target and, in the presence of shocks, the inflationary and deflationary policies

that may be needed to maintain the price target. This point becomes important since the transparency and credibility of policy is important to the public's expectations on inflation and the impact of monetary policy on the economy. Therefore, central banks have tended to adopt inflation targeting rather than price-level targeting. The popular Taylor rule embodies this preference for inflation rather than price-level targeting, with the target inflation rate that is usually set for developed economies being in the range from 1 percent to 3 percent.

10.6 Determination of the money supply

No matter how the money supply in the economy is defined or measured, several major participants are involved in its determination. They are:

1 The central bank, which, among its other policies, determines the monetary base and the reserve requirements for the commercial banks, and sets its discount rate.
2 The public, which determines its currency holdings relative to its demand deposits.
3 The commercial banks, which, for a given required reserve ratio, determine their actual demand for reserves as against their demand deposit liabilities.[11]

Some indication of the relative importance of the major contributors to changes in the money supply would be useful at this point. Phillip Cagan (1965) concluded that, in the USA, on average over the 18 cycles during 1877 to 1954, the fluctuations in the currency ratio had a relatively large amplitude over the business cycle. They caused about half of the fluctuations in the growth rate of the money stock, while fluctuations in the monetary base and the reserve ratio accounted for roughly one-quarter each. But, from a secular perspective, by far the major cause of the long-term growth of the money stock was the growth in the monetary base.

Therefore, there is considerable interaction between the behavior of the central bank, the public and the commercial banks in the money supply process. This interaction is important in studying the behavior of the central bank which, as a policy-making body deciding on the total amount of money desirable for the economy, must take into account the responses of the public and of the commercial banks to its own actions. The behavior of the central bank in the money-supply process becomes a distinctive topic of study, which is pursued later in this chapter and the next two chapters.

10.6.1 Demand for currency by the public

Fluctuations in the public's demand for currency relative to its holdings of demand deposits are a significant source of fluctuations in the money supply. The closest substitute – and a fairly close one at that – to currency holdings (C) is demand deposits (D), so that most studies on the issue examine the determinants of the ratio C/D, or of the ratio of currency to the total money stock (C/M1), rather than directly the determinants of the demand for currency alone.

The C/D ratio varies considerably, with a procyclical pattern over the business cycle and over the long term. The desired C/D ratio depends upon the individual's preferences in the

11 These are not the only actors in the money supply process. In particular, in open economies, the balance of payments surpluses (deficits) of a country can increase (decrease) its money supply.

light of the costs and benefits of holding currency relative to demand deposits. Some of these costs and benefits are non-monetary and some are monetary.

The non-monetary benefits and costs are related to the non-monetary costs of holding and carrying currency compared with those of holding demand deposits and carrying checks. They also take into account the general acceptability of coins and notes for making payments as against payments by other means. In financially less developed economies, with few bank branches in the rural areas and with banking usually not open to or economically feasible for lower income groups, even in the urban areas, currency has a clear advantage over checks. However, even in financially advanced economies, cash is almost always accepted for smaller amounts while the use of checks is restricted to payments where the issuer's credit-worthiness can be established, or the delivery of goods can be delayed until after the clearance of the check through the banks. It is also more convenient to make very small payments in cash than by writing a check. These aspects of non-monetary costs have changed substantially over time in favor of bank deposits with the expansion of the banking system and the modernization of its procedures, increasing urbanization, spread of banking machines, common usage of credit and debit cards, etc.

As against the greater convenience of currency over bank deposits for transactions, the possession of a significant amount of currency involves risks of theft and robbery, which impose not only its loss but also a risk of injury and trauma to the carrier. The fear of the latter is often sufficient to deter possession of large amounts of currency in societies where this kind of risk is significant. This is so in most countries, with the result that only small amounts of currency are carried by most individuals at one time or stored in their homes. By comparison, the demand for currency in Japan – a society with a very low rate of theft and robbery – is dominated by its convenience relative to bank deposits. Consequently, few individuals in Japan hold demand deposit accounts, checks are not widely accepted in exchange, even by firms, or given by them for payment of salaries. Many transactions, even of fairly large amounts, are done in currency.

The monetary costs and benefits of holding currency relative to demand deposits really relate to the net nominal return on the latter since currency does not have an explicit monetary return or service charge, while demand deposits often possess one or both of these. In any case, even if demand deposits pay interest, there is usually a negative return on them since banks incur labor and capital costs in servicing them and must recoup these through a net charge on them.[12]

Chapter 4 presented the inventory analysis of the demand for money. This model is applicable to the demand for currency relative to demand deposits. This was done in Chapter 4. As pointed out there, the problem with an application of this analysis that takes account only of the monetary cost of using currency versus demand deposits is that it ignores the non-monetary differences in their usage: acceptance in certain types of payments, risks of theft and robbery, etc. However, its central conclusion still holds: the optimal holdings of currency relative to demand deposits will depend on their relative costs and the amount of expenditures financed by them. Therefore, assuming both currency and demand deposits are "normal goods," an increase in the net cost of

12 One estimate of the average total cost of demand deposits in 1970 was about 2.4 percent of their dollar volume. The cost of time deposits was, by comparison, 0.6 percent if the interest costs were excluded and between 5.3 percent and 5.7 percent if the interest costs were included. Approximately 70 percent of the cost associated with demand deposits was the cost of processing checks and involved wages, computer time costs, etc.

holding demand deposits would increase the demand for currency and hence the C/D ratio.

However, in a time-series context, the major reasons for changes in this ratio have been the innovations in shopping, payments and banking practices which have made checking increasingly easier and thereby lowered the C/D ratio. In addition, the significant possibility of theft and robbery – and the consequent risk to the person – with increases over time in many economies, have kept this ratio quite low or further reduced it. As indicated earlier, Japan, with a low risk from such criminal activities, is an exception to this rule and illustrates the greater convenience of using currency where a sufficiently wide range of denominations is made available in bank notes.

For the future, in financially developed economies, smart cards are likely to become a close substitute for currency in many transactions that used to be settled in currency since such cards may prove to be even more convenient than currency and yet not more susceptible to theft. Therefore, the demand for currency as a proportion of total expenditures or of M1 or M2 is likely to continue to decline in the future.

The above arguments imply that the demand function for currency can be written as:

$$\mathrm{C/D} = c(\gamma_{\mathrm{D}}, R^{\mathrm{h}}, R_{\mathrm{D}}, R_{\mathrm{T}}, Y; \text{payments technology}) \tag{4}$$

where:

c = currency-demand deposit ratio
γ_{D} = charges on demand deposits
R^{h} = average yield on the public's investments in bonds, etc.
R_{D} = interest rate on demand deposits
R_{T} = interest rate on time deposits
Y = nominal national income

$\partial c/\partial \gamma_{\mathrm{D}} > 0$ and $\partial c/\partial R_{\mathrm{D}} < 0$ for obvious reasons. We expect $\partial c/\partial Y > 0$, since an increase in Y increases transactions that are likely to increase the demand for currency proportionately more than for demand deposits. This implies that the currency ratio will increase in the upturns and decrease in the recessions. An increase in the return on both time deposits and bonds is likely to decrease the demand for both currency and demand deposits. Further, currency is needed for small everyday transactions that tend to be inelastic in response to changes in interest rates, while efficient cash management techniques allow reductions in demand deposits. Hence, the currency ratio will rise with increases in R^{h} and R_{T}, so that $\partial c/\partial R^{\mathrm{h}} > 0$ and $\partial c/\partial T_{\mathrm{T}} > 0$. This implies that an increase in the rate of inflation and/or in the nominal interest rate, as usually happens in the upturn of the business cycle, would increase the currency ratio. Conversely, this ratio will fall in a recession. Hence, we expect the currency ratio to be procyclical (i.e. to rise in upturns and fall in downturns).

As stressed above, the currency ratio also depends upon the security environment and the availability of alternative modes of payment such as debit and credit cards. The impending innovations in creating smart cards that represent electronic purses are likely to reduce currency demand.

Households not only hold currency and demand deposits but also hold various forms of savings deposits, term deposits and their variants. All of these pay interest, and we can specify the arguments that would lead to the public's demand function for time deposits or for its desired ratio of time to demand deposits. The derivation of these functions is left to the reader.

10.6.2 Commercial banks: the demand for reserves

Commercial banks hold reserves against their deposits. A part of these reserves is normally held in cash (at the tills, in the automatic teller machines or in the bank's vault) and part is held in deposits with the central bank. If only a small part of deposits is withdrawn from a bank during a period, the bank does not have to maintain reserves equal to deposits (i.e. follow a 100 percent reserve ratio) but could increase its revenues by lending out a part or most of its deposits. This leads to *fractional reserve banking*, where the fraction of deposits held in reserves may be quite small, as discussed later in this chapter. The *reserve ratio* is the ratio of reserves held to deposits.

The central bank often requires the commercial banks to meet a certain minimum ratio called the *required reserve ratio*[13] – of their reserves to their deposit liabilities.[14] Chapter 11 will present the required reserve ratios for several countries. In 1999, this ratio was zero in Canada and the United Kingdom. In the United States, it depended on the amount of deposits and varied between 3 percent and 9 percent for depository institutions.

Banks usually hold reserves in excess of those required to meet the required reserve ratio. Banks also borrow from other banks or the central bank. Reserves held in excess of the sum of required and borrowed reserves are referred to as *free reserves* – that is, at the bank's disposal for use if it so desires.

Free reserve hypothesis

Free reserves for a bank are those it wants to hold in addition to its required reserves and borrowed reserves. Free reserves must be distinguished from *excess reserves*, which are actual holdings of cash reserves in excess of the sum of required, borrowed and free reserves. Excess reserves are ones that the bank wants to eliminate either immediately or gradually. The hypothesis for the determination of free reserves is known as the *free reserve hypothesis*.

Required reserves and free reserves depend upon the required reserve ratio or differential ratios imposed by the central bank, discussed in Chapters 11 and 12 on central bank behavior, and upon the total deposits in the bank. The computation of such required reserves is largely mechanical, according to a formula prescribed by the central bank.[15]

Each bank has to anticipate its deposit liabilities in making its decisions on its reserve holdings. Demand deposits may be withdrawn on demand at any time and individuals' demand deposits in any given bank fluctuate considerably over time as they make deposits and withdrawals. For any given bank over a given period, the totals of new deposits and withdrawals are likely to cancel out to some extent, depending upon its size and the distribution of its depositors among occupations and industries, between employees and employers, etc. The degree of uncertainty as to the average levels of deposits in any bank is

13 The next chapter on the central bank discusses required reserve ratios in greater detail. It is now close to zero in many Western countries, excluding the United States where it is between 3 percent and 10 percent.

14 Some countries, including the United States, have different required ratios for different types of bank liabilities or for different financial institutions, depending on their size and nature.

15 The formulae for computing legally required reserves differ between countries, with significant impact on the behavior of banks toward free reserves and hence toward monetary policy. A significant difference can be whether the reserves required to be held by a bank are computed as a proportion of its deposits in the current period or some past period, say a few weeks earlier. The former process introduces uncertainty in the amount of required reserves, while the latter does not.

thus likely to vary between banks. It is likely to be higher for unit rather than branch banks, small rather than large banks, and banks with a smaller degree of monopoly than those with a greater one. The cancellation process is likely to be still greater for all the banks in the economy taken together, so that the overall amounts of demand deposits in the economy normally exhibit a great deal of stability.

For banks, reserves are an asset, in addition to bonds and loans, so that the demand for reserves depends on the returns on the latter. Reserves do not generally earn a monetary return. Their demand should, because of the substitution effect, fall as the rates of return on the other assets rise, and vice versa.

Under uncertainty, the free reserve hypothesis assumes that banks maximize the expected utility of their terminal wealth, corresponding to the expected utility maximization by the individual investor presented in Chapter 5. Therefore, the theory of portfolio selection set out in Chapter 5 can be adapted to explain the bank's demand for free reserves. Assuming that the banks are risk averters, disliking the prospect of ending up with less than the required reserves, they would always hold more than the required reserves. Part of these extra reserves could be borrowed, so that the demand for free reserves will depend upon the risks present, the response to risk, the cost of borrowing and the return on other assets in the bank's portfolio.

A significant element of the risk in falling short of the desired reserves depends on the structure of the banking system, the size of the bank in question and the diversity of its client base. Canadian and British banks tend to be large, with branches all over the country. They have a very diversified client base, so that the daily variance in their deposits is relatively low. The US banks in the smaller cities and rural areas are often small, have a limited number of branches and may be dependent on a particular segment of the economy. Consequently, they face higher daily variance in their deposits.

Another significant element of risk in falling short of the desired reserves is related to the formula specified by the central bank for the minimum reserves that the banks should hold against their deposits.[16] In the UK, although the reserve requirement is zero (or, rather, non-negative), the banks are expected to meet it on a daily basis. This increases their risk, which to some extent is offset through their ability to borrow reserves in the overnight market from other financial institutions. Canada allows averaging of reserves and deposits over a 4 to 5-week period and the United States allows this over 2 weeks, thereby implying less risk for their banks than for British banks.[17]

Borrowing by commercial banks from the central bank

Banks borrow reserves from a variety of sources. Banks frequently borrow reserves from each other bilaterally and often do so in the context of an organized overnight loan market such as the Federal Funds market in the USA and the Overnight Loan market in Canada. They can, depending on regulations, also borrow abroad to supplement their reserves. Borrowing by individual banks from other banks within the system does not change the monetary base and can therefore be ignored in the determination of the money supply. However, when the

16 Information on these is given in Chapter 11, Table 11.1.

17 Another consideration is whether lagged reserve accounting is allowed, i.e. whether the banks can average their reserves and deposits over a specified past period or have to include the current one. The former involves less risk.

commercial banks as a whole increase their borrowing from the central bank or from abroad, the monetary base increases and the money supply expands.

In lending to the commercial banks as a whole, the central bank is said to act as the lender of last resort since the banking system as a whole cannot obtain additional funds from its own internal borrowing and lending. However, individual commercial banks sometimes treat the central bank as their lender of first resort rather than last resort. The terms on which it lends and the conditions under which its loans are made affect the amounts that the banks wish to borrow from it. In general, borrowing from the central bank can trigger greater oversight by the central bank into the borrowing bank's asset management and other practices. Since this is rarely desired, it acts as a disincentive to borrow.

In the USA, the discount rate – at which the Federal Reserve System lends to its member banks – is usually below the three-month Treasury bill rate, so that these banks stand to gain by borrowing from the Federal Reserve System. To limit the amounts and frequency of borrowing, the Federal Reserve Board imposes a variety of formal and informal rules on such borrowing. One of the latter is that borrowing from the Federal Reserve System is a *privilege*, extended by the Federal Reserve System to its member banks, rather than a right of the banks. This privilege can be curtailed or circumscribed by conditions if a bank tries to use it indiscriminately.

Canada has experimented with two different methods for setting the bank rate at which it lends to the chartered banks. Under a fixed bank rate regime adopted from 1956 to 1962 and from 1980 to 1994, the bank rate was automatically set each week at 0.25 percent higher than the average Treasury bill rate that week. Since it was higher than the Treasury bill rate, the chartered banks incurred a loss if they financed their purchases of Treasury bills by borrowing from the Bank of Canada. Such a rate is said to be a "penalty rate" and, by its nature, discourages borrowings. It does not therefore need the support of other restrictions on borrowings to the extent that the American discount rate requires. Since 1994, the Bank has placed its major focus on setting the overnight loan rate, with an operating band around it of 50 basis points. This rate is the rate at which the banks and other major participants in the money market make overnight loans to each other. Since 1996, the bank rate has been set at the upper limit of the operating range for the overnight loans, so that it is a floating rate and continues to be a penalty rate, irrespective of daily movements in the market rates. The Bank of Canada influences the Bank Rate by changes in its supply of funds to the overnight market or through its purchases or sales of Treasury bills.

In the United Kingdom, the Bank of England determines daily the rate at which it will lend to the banks. This allows it control over borrowings from it on a daily basis. It also allows close control over the interest rates that the banks charge their customers, since these rates have base rates closely tied to the Bank's daily rate.

Banks' demand function for reserves

The free reserve hypothesis, in attempting to explain the demand for free reserves, has to take these differing practices into account and would yield differing demand functions for different countries and periods. However, empirical studies[18] have confirmed its implication that the earnings on alternative assets influence the amount of reserves held and that the ratio

18 For example, see DeLeeuw (1965), Goldfeld (1966), Frost (1971).

of desired reserves to demand deposits cannot be taken as constant for purposes of monetary policy.

The preceding arguments imply that the demand function for desired free reserves, FR, can be expressed as:

$$\mathrm{FR}/D = f(R, R_{\mathrm{BR}}, R_{\mathrm{CB}}) \tag{5}$$

where:

R = average interest rate on banks' assets
R_{BR} = average return on banks' reserves
R_{CB} = central bank's discount rate (for loans to the commercial banks)

In (5), $\partial f/\partial R < 0$, $\partial f/\partial R_{\mathrm{BR}} > 0$ and $\partial f/\partial R_{\mathrm{CB}} > 0$. We have simplified this function by including average interest rates rather than the variety of interest rates that will need to be considered in practice.

Note that, in (5), R is an average of the nominal returns on Treasury bills, bonds of different maturities, mortgages and loans to the public, etc. Also note that R_{BR} would be zero for reserves held in currency and would also be zero if the rest of the reserves were held in non-interest paying deposits with the central bank. Banks will want to increase free reserves if the cost R_{CB} of covering a shortfall in reserves increases and decrease them if the return R from investing their funds rises. Free reserves would also increase if R_{BR} were positive and were to increase.

10.7 Mechanical theories of the money supply: money supply identities

Mechanical theories of the money supply are so called because they use identities, rather than behavioral functions, to calculate the money supply. The money-supply equations resulting from such an approach can be easily made more or less complex, depending upon the purpose of the analysis. We specify below several such equations, starting with the most elementary one.

An elementary demand deposit equation

Assume that the ratio of reserves held by the banks against demand deposits is given by:

$$BR = \rho D$$

where:

BR = reserves held by banks
D = demand deposits in banks
ρ = reserve ratio

If ρ equals the required reserve ratio, set by the central bank for the banking system, and BR represents the reserves exogenously supplied to it, profit-maximizing banks will create the amount of deposits given by:

$$D = (1/\rho)BR \tag{6}$$

This equation is the elementary deposit creation formula for the creation of deposits by banks on the basis of the reserves held by them. It suffers from a failure to take note of the behavior of the banks and the public in the deposit expansion process, so that a more elaborate approach to the money supply is preferable.

Common money-supply formulae

Friedman and Schwartz (1963) used a money-supply equation that not only takes account of the reserve/deposit ratio of banks but also of the ratio of currency to deposits desired by the public. Their ratio is derived simply from the accounting identities:

$$M = C + D \tag{7}$$

$$\text{M0} = BR + C \tag{8}$$

where:

D = demand deposits
BR = banks' reserves
C = currency in the hands of the public
M0 = monetary base = $BR + C$.

The Friedman and Schwartz money-supply formula is derived as follows:

$$\begin{aligned} M &= \frac{M}{\text{M0}}\text{M0} \\ &= \frac{C+D}{BR+C}\text{M0} \\ &= \frac{1+D/C}{BR/C+1}\text{M0} \\ &= \frac{(1+D/C)(D/BR)}{(BR/C+1)(D/BR)}\text{M0} \\ &= \frac{(1+D/C)(D/BR)}{(D/BR+D/C)}\text{M0} \end{aligned} \tag{9}$$

Equation (9) separates the basic determinants of the money stock into changes in the monetary base and changes in the "*monetary base multiplier*," defined as ($\partial M/\partial$M0), for the monetary base.[19] This multiplier is itself determined by D/BR, the reserve ratio, and C/D, the currency ratio. Of these, the reserve ratio reflects the required reserve ratio and the banks' demand for free reserves. The currency ratio reflects the public's behavior in its demand for currency. Hence, the three determinants of the money stock emphasized by (9) are the monetary base, the currency and the reserve ratios.

Another version of the money-supply formula is:

$$M = \frac{1}{\left(\frac{C}{M} + \frac{BR}{D} - \frac{C}{M}.\frac{BR}{D}\right)}\text{M0} \tag{10}$$

Using this equation, Cagan (1965) examined the contributions of the three elements B, C/M and BR/D, to M2 over the business cycle and in the long term. He found that the dominant

19 We have designated $\partial M/\partial$M0 the *monetary base multiplier*, so that it captures the impact of changes in the monetary base on the money supply. Some other authors call it the *money multiplier*. We have left the latter concept of multiplier to mean $\partial Y/\partial M$ (where Y is national income), as used in Chapters 13 to 15, on the macroeconomy.

factor influencing the long-term growth in the money stock was the growth in the monetary base. Changes in the two ratios contributed little to the *secular* change in M2. However, for *cyclical* movements in the money stock the changes in the C/M ratio were the most important element, whereas the reserve ratio had only a minor impact and changes in the monetary base exerted only an irregular influence.

The currency–demand deposit – and hence the currency–money ratio – is influenced strongly by changes in economic activity and especially by changes in the rate of consumer spending. As we have explained in earlier sections, this ratio varies in the same direction as nominal national income – hence, pro-cyclically – so that a rise in spending in cyclical upturns increases currency holdings, which lowers the money supply.

The preceding money-supply formula does not differentiate deposits into various types such as demand deposits, time and savings deposits, and government deposits nor does it differentiate between their reserve requirements. A formula (its derivation is not shown) that does so is:

$$M = \left[\frac{1+c}{\rho_D + \rho_T t + \rho_G g + c}\right] \mathrm{M0} \tag{11}$$

where $t = T/D$ and $g = G/D$, T and G represent time/savings deposits and government deposits respectively in commercial banks.

The above formulae are all identities. Which one is used depends upon the rules and regulations about reserve ratios, the availability of statistical data and the further behavioral assumptions that are made. In practice, theories of the money supply go beyond these identities and embed the relevant identity in a behavioral theory.

10.8 Behavioral theories of the money supply

A behavioral theory of the money supply process must take into account the behavior of the different participants in this process in order to determine the economic and non-economic determinants of the variables being studied. Such a theory studies this behavior in terms of the main components of the preceding money supply formulae, such as the currency desired by the public, the reserves desired by the commercial banks, the amounts borrowed by them and the monetary base which the central bank wishes to provide.

Our earlier discussions on the demand for components of the money supply imply the general form of the money-supply function to be:

$$M^{\mathrm{s}} = M^{\mathrm{s}}(R_D, R_T, R_S, R_L, R_d, R_O, R, Y, \mathrm{M0}) \tag{12}$$

where:

R_D = charges on demand deposits
R_T = interest rate on time deposits
R_S = short-term interest rate
R_L = long-term interest rate
R_d = discount rate (central bank rate for lending to the commercial banks)
R_O = overnight loan rate
RR = required reserve ratio.

As argued in earlier sections, the money supply depends, among other variables, upon the free reserves desired by the banks and the currency desired by the public. Free reserves depend

upon R_O, R_d, R_S and R_L since these determine the opportunity cost of holding free reserves. The public's demand for currency will similarly be a function of R_{D} and R_{T}. It also depends, as argued earlier, on the level of economic activity for which nominal national income Y is a proxy. Finally, the money supply depends upon the monetary base M0.

The monetary base is under the control of the monetary authorities, which can operate it in such a way as to offset the effect of changes in the other variables on the money supply. Alternatively, they may only allow the changes in the other variables to affect the money supply in so far as the effect changes the money supply to a desired extent. The monetary base is, therefore, not necessarily a variable independent of the other explanatory variables in (12).

Now consider the directions of the effects that are likely to occur in the money-supply function. An increase in the monetary base increases the money supply. An increase in national income increases the currency demand and lowers banks' reserves and hence decreases the money supply. An increase in the short-term market interest rate increases the profitability of assets which are close substitutes for free reserves and hence decreases the demand for free reserves, which increases the money supply. A cut in the discount rate has a somewhat similar effect. An increase in the yield on time deposits increases their demand by the public, which lowers the reserves available for demand deposits, so that demand deposits decrease.

In practice, the estimation of the money-supply function is usually undertaken with a smaller number of variables than those specified in (12). This is partly because of collinearity among the various interest rates.

Equation (12) specifies the money-supply function and could be applied to the supply of either M1 or M2 or another monetary aggregate. However, note that the signs of the interest-rate elasticities could differ among the aggregates. There are three main cases of differences:

1 If the interest rates on demand deposits increase, their demand will increase but this could be merely at the expense of time deposits, so that while M1 increases, M2 does not do so. Hence, the interest elasticity of demand deposits and M1 with respect to R_D is positive but that of M2 may be positive or zero.
2 Since time deposits are excluded from M1, they are part of the opportunity cost of holding M1. Hence, the elasticity of M1 with respect to the interest rate on time deposits, R_T, would be negative. These deposits are, however, part of M2 so that, when their interest rate increases, the desired amount of time deposits increases and so does M2. That is, the interest elasticity of M2 with respect to R_T is positive.
3 The interest elasticity of M1 would be negative with respect to the return on bonds. But if the time deposits interest rates rise with this bond rate, the respective interest elasticity of M2 is likely to be zero. However, if the time deposit rates do not increase when the bond rate rises, the public will switch some time deposits to bonds, so that the elasticity of M2 with respect to R would be negative. Therefore, this interest elasticity depends upon the relationship between R_T and R.

Table 10.2 shows the pattern of interest-rate elasticities for M1 and M2. This table also includes the elasticities of M1 and M2 with respect to the return on excess reserves and the central bank discount rate on borrowed reserves. Both are negative. The reasons for these and the other signs shown in Table 10.2 have been explained above.

Table 10.2 Money supply elasticities

Interest rate on:	*M1*	*M2*
Demand deposits	+	?
Time deposits	−	+
Bonds	−	?
Excess reserves	−	−
Central bank's discount rate	−	−

Interest rate elasticities of the money supply

Empirical studies on money supply are far fewer than on money demand. The following brief review of the empirical findings on the money supply function confines itself to reporting the elasticities' estimates for this function.

Rasche (1972) reports the impact and equilibrium elasticities calculated by him for the supply functions reported by DeLeeuw (1965) for the Brookings model, by Goldfeld (1966) for the Goldfeld model, and for the MPS model developed by the Federal Reserve–MIT–Pennsylvania econometric model project. These studies were for the USA, using data up to the mid-1960s. Rasche's calculations of these elasticities are roughly in the ranges reported in Table 10.3.[20]

Note that there have been many changes in the United States' financial markets since the 1960s, so that the elasticity ranges reported in Table 10.3 are now mainly useful for pedagogical purposes. These elasticities indicated that the main components of the money supply – and the money supply itself – were not exogenous but functions of the interest rates in the economy. We conclude from Table 10.3 that:

1 Currency demand was negatively related to the time deposit rate, whereas time deposits were positively related.
2 Time deposit holdings were positively related to their own interest rate and negatively to the Treasury bill rate.
3 Banks increased their borrowing from the Federal Reserve as the Treasury bill rate rose and decreased them as the discount rate increased. Note that the discount rate is the cost of such borrowing and the Treasury bill rate represents the return on funds invested by the banks. Hence an increase in the Treasury bill rate provides an incentive for banks to increase their loans for a given monetary base, and also to increase their borrowing, given the discount rate, from the central bank. Both these factors imply a positive elasticity of the money supply with respect to the Treasury bill rate, as shown in Table 10.3.
4 Conversely, while the banks' free reserves decreased as the Treasury bill increased, they increased with the discount rate. The Treasury bill rate represents the amount the banks lose by holding free reserves and is, therefore, their opportunity cost, so that the free reserves fall with the Treasury bill rate. However, these elasticities with respect to the discount rate are positive since the discount rate is the "return" on free reserves; i.e. if they hold adequate free reserves to meet withdrawals, the banks escape having to borrow from the central bank at the discount rate.

20 These estimates cover different studies and different periods, and are reported for illustrative purposes only.

Table 10.3[a] Interest rate elasticities for the money supply function

	Impact[b]	*Equilibrium*
Currency demand		
Time deposit rate	−0.012 to −0.015	−0.136 to −0.14
Treasury bill rate	−0.008 to 0.0037	−0.07 to 0.026
Time deposit demand		
Treasury bill rate	−0.038 to −0.15	−0.374 to −1.4
Time deposit rate	0.070 to 0.3	0.683 to 2.9
Bank borrowings		
Treasury bill rate	0.134 to 0.88	0.50 to 2.625
Discount rate	−0.186 to −0.98	−0.70 to −2.926
Free reserves		
Treasury bill rate	−2.99 to −3.95	−6.42 to −8.47
Discount rate	3.23 to 3.48	6.93 to 7.46
Money supply, excluding time deposits[c]		
Treasury bill rate	0.214 (average over first 6 months) 0.267 (average over second 6 months) 0.2438 (average over third 6 months)	
Money supply, including time deposits[d]		
Treasury bill rate	0.219 (average over first 6 months) 0.278 (average over second 6 months) 0.258 (average over third 6 months)	

a Compiled from Rasche (1972), various tables.
b The impact elasticity is over the first quarter following a change in the interest rate.
c From Rasche (1972), Table 5.
d From Rasche (1972), Table 5.

These reported elasticities are consistent with the analysis presented earlier in this chapter. Even though there have been numerous innovations in the financial markets since the 1960s, so that the magnitudes of the actual elasticities are likely to have changed, there is no reason to expect that the *signs* of elasticities have altered from those reported above.

Lags in the money-supply function

The main findings on this show that:

1 The impact elasticities are significantly lower than the equilibrium ones, indicating that the adjustments take longer than one quarter.
2 The money supply had positive interest elasticities each month through the first 18 months for which the elasticities were reported.

These findings show that the financial sector did not adjust the money supply to its full equilibrium level within one quarter. In fact, the second finding points out that the money supply continues to change even after six quarters. This finding has been confirmed by many studies, so that the existence of lags in the response of the money supply to interest-rate changes can be taken as being well established.

10.9 Cointegration and error-correction models of the money supply

There are few cointegration studies on the money supply function and its major components. We draw the following findings from Baghestani and Mott (1997) to illustrate the nature of empirical findings on money supply and the problems with estimating this function when monetary policy shifts.

Baghestani and Mott performed cointegration tests on USA monthly data for three periods, 1971:04 to 1979:09, 1979:10 to 1982:09 and 1983:01 to 1990:06, using the Engle–Granger techniques. Their variables were log of M1, log of the monetary base (B) and an interest rate variable (R). The last was measured by the three-month commercial paper rate for the first two periods and by the differential between this rate and the deposit rate paid on Super NOWs (Negotiable Orders of Withdrawal at banks) introduced in January 1983. Further, the discount rate was used as a deterministic trend variable, since it is constant over long periods. The data for the three periods was separated since the Federal Reserve changed its operating procedures between these periods.

Baghestani and Mott could not reject the null hypothesis of no cointegration among the designated variables for 1971:04 to 1979:09. Further, for 1979:10 to 1982:09, while M0 and R possessed a unit root, M1 did not, so the cointegration technique was not applied for this period. The only period which satisfied the requirement for cointegration and yielded a cointegration vector was 1983:01 to 1990:06. The error-correction model was also estimated for this period. The cointegration between the variables broke down when the period was extended beyond 1990:06. These results have to be treated with great caution. As indicated in Chapter 9 on money-demand estimation, cointegration is meant to reveal the long-run relationships and, for reliable results, requires data over a long period rather than more frequent observations, as in monthly data, over a few years. The three periods used by Baghestani and Mott were each less than a decade.

For 1983:01 to 1990:06, Baghestani and Mott concluded from their cointegration–ECM results that the economy's adjustments to the long-run relationship occurred through changes in the money supply and the interest rate, rather than in the monetary base. Comparing their findings across their three periods, we see that changes in the central bank policy regime, such as targeting monetary aggregates or interest rates, are extremely important in determining the money supply function in terms of both its coefficients and whether there even exists a long-run relationship. Further, even regulatory changes such as permitting, after 1980, the payment of interest on checkable deposits can shift the money-supply function.

10.10 Monetary base and interest rates as alternative policy instruments

The central bank may use either the monetary base or the interest rate as a way of controlling aggregate demand in the economy, or may have to use both. Under certainty and known money supply and demand functions, it needs to use only one of them since the use of either of them indirectly amounts to use of the other. To see this correspondence, assume that the money supply function is given by:

$$M = \frac{\mathrm{M0}}{\left[\frac{C}{M} + \frac{BR}{D} - \frac{C}{M} \cdot \frac{BR}{D}\right]} \tag{13}$$

where the meanings of the symbols are as explained earlier. This money supply function can be written simply as:

$$M = \alpha \mathrm{M0}, \tag{14}$$

where:

$$\alpha = \frac{1}{\left[\frac{C}{M} + \frac{BR}{D} - \frac{C}{M}.\frac{BR}{D}\right]}$$

where α is the "monetary base (to money supply) multiplier" $\partial M/\partial \mathrm{M0}$, though some authors call it the "money multiplier," a term that we have more appropriately reserved for $\partial Y/\partial M$, where Y is nominal national income.

Let the general form of the money demand function be:

$$m^{\mathrm{d}} = m^{\mathrm{d}}(y, R) \tag{15}$$

where R is the nominal interest rate and y can be the pre-specified actual or desired level of output. In money market equilibrium, we have:

$$\alpha.\mathrm{M0} = P.m^{\mathrm{d}}(y, R) \tag{16}$$

Under certainty, given the policy targets for P at P^* and y at y^*, (16) can be solved for the relationship between M0 and R, so that the central bank can achieve its objectives by setting the monetary base M0 at M0* and letting the economy determine R, or by setting R at R^* and letting the economy determine the money supply needed to support R^*. It does not have to pursue a policy of setting both M0 and R.

Making the choice between the interest rate and the monetary base as operating targets in a stochastic context

In developed economies, adequate reasons for the choice between M0 and R as the optimal monetary policy instrument arise only if there is uncertainty and unpredictability of the money supply and/or demand functions.[21] In this scenario, the policy maker may not know which policy instrument would more predictably deliver the target values y^* and P^*. In general, in the absence of any sure information, the theory of policy under uncertainty implies that the risk-averting policy maker should diversify by using both policy instruments. However, as Poole's analysis presented earlier in this chapter (and in Chapter 13) shows, if the main shocks to aggregate demand originate from shifts in the commodity sector, then the money supply is, in general, the preferable monetary policy instrument. But, if the main shocks to aggregate demand originate in the monetary sector, the interest rate is, in general, the preferable one (Poole, 1970). The next chapter examines this issue in greater detail.

21 Note that even if the deterministic elements of the two functions are known, there will also be a stochastic component which will be unpredictable.

Another reason for the choice between the two instruments can arise because of the nature of the economy. Underdeveloped financial economies usually do not have well-developed bond markets, which prevents the central bank from effectively using open-market operations. However, there may be other ways of changing the monetary base, such as by feeding increases in the money supply to the government to finance its deficits. In addition, changes in the reserve requirements can be used to change the supply for a given monetary base. Such economies also have fragmented financial markets, along with a large informal financial sector, so that there need not be a close relationship between the interest rate set by the central bank and that charged in the various private financial markets. Further, much of the investment in these economies may not be sensitive to market interest rates. Therefore, on the whole for such economies, it is likely that changes in the money supply will be more effective, though not perfectly so, in manipulating aggregate demand than will changes in the interest rate. However, given the imperfections of both policy instruments in controlling aggregate demand, demand management would work better if the central bank were to use both instruments.

Conclusions

This chapter has examined the implications of adopting different operational targets for monetary policy. For most central banks, the direct target is no longer a reserve aggregate such as borrowed or non-borrowed reserves, the monetary base or a monetary aggregate (M1, M2, M3 or M4), since these proved to have an unstable relationship with nominal GDP in the 1980s and 1990s. The main reason for this instability has been financial innovation leading to changes in the velocity of circulation, new monetary assets and increased substitutability between demand deposits and other financial assets.

Even though the role of the money supply has been downgraded in recent years, it remains critical to the economy's economic performance. The theory of money supply for any given economy therefore, has to start with the knowledge of whether or not its central bank exogenously determines the money supply, thereby canceling out any undesired changes in it induced by other factors in the economy. If it does so, the final determinant of the money supply is central bank behavior. However, if it allows other factors some role in changing the money supply, a wider analysis of the money supply becomes applicable and the behavior of the public and of financial institutions also needs to be studied.

The theories of money supply can be specified as either mechanical or behavioral. In practice, most empirical studies combine both approaches in deriving their estimating equations. These studies show that, over time, changes in the monetary base are the main element of changes in the money supply. However, over the business cycle, changes in the currency and the reserve ratios also play a very significant role, with the result that their determinants also need to be incorporated into the estimating equations. These determinants include a variety of interest rates in the economy, as well as national income. The estimated money supply functions, especially for M1, usually show negative elasticities with respect to the central bank discount rate and positive elasticities with respect to the Treasury bill rate.

The application of cointegration techniques to the money supply and its determinants is not as common as for the money demand function. Cointegration requires long runs of data for reliable results. Since the money supply depends critically upon central bank behavior, periodic changes in the central bank's targets and money supply rules make it difficult

to collect sufficient data for reliable cointegration results on the money supply function. However, the reported empirical studies on this function do usually yield signs consistent with the signs of the interest-rate elasticities implied by the theory.

Summary of critical conclusions

- Successful interest-rate targeting, in comparison with monetary targeting, increases the impact on aggregate demand of investment, net exports, fiscal deficits and other disturbances in the commodity markets while eliminating the impact of shocks emanating from the financial sectors.
- Monetary targeting eliminates the impact of fluctuations in the money supply induced by the private sector and moderates the impact of fluctuations emanating from the commodity market.
- The public and the banks affect the money supply by changes in the currency ratio and free reserves, so that the central bank must be able to predict these ratios if it is to offset them and control the money supply in the economy.
- Empirical studies support a behavioral approach to the money supply and indicate that the money supply depends on the interest rates in the economy. The estimated money supply functions are sensitive to monetary policy and target regimes.

Review and discussion questions

1. Discuss, using diagrams, the following statement: if the money demand function has a high interest inelasticity, the case for a monetary aggregate as against an interest rate as the central bank's operating target is strengthened, especially in an economy subject to stochastic shocks.
2. How do central banks manage interest rates in your country and one other country of your choice? What consequences for output fluctuations can the central bank expect from targeting interest rates?
3. The monetary sector has become increasingly unstable in recent years. Does this mean that the monetary authority should stay with the pursuit of interest-rate targets and leave the money supply alone?
4. Note that recessions seem to be caused by either reductions in aggregate demand or in aggregate supply, or by the two acting in concert. What targets should the central bank adopt? Would the optimal choice of the target be the same for reductions in aggregate supply as for reductions in aggregate demand?
5. "Macroeconomic models assume that the money supply is exogenously specified by the central bank. If this is so, there is no purpose to the specification and estimation of money supply functions." Discuss this statement. How would you verify its validity?
6. What happens to the monetary base and the money supply if the government finances a fiscal deficit by:

   (a) selling bonds to the public;
   (b) selling bonds to the commercial banks;
   (c) selling bonds to the central bank;
   (d) selling bonds to foreigners.

If any of these changes the money supply, what policies should the central bank adopt to offset these changes?

7. What happens to the monetary base and the money supply if:

   (a) the central bank lowers the discount rate;
   (b) the central bank lowers the discount rate and also sells bonds to the public;
   (c) the central bank forbids overnight loans and eliminates the overnight loan market.

8. Show what happens to the money supply if:

   (a) the economy enters a boom and interest rates rise;
   (b) the underground economy with illegal holdings of currency is eliminated;
   (c) firms give a significant discount for payment in cash rather than credit cards;
   (d) credit cards are replaced totally by debit cards;
   (e) both credit and debit cards are replaced by smart cards.

9. Does the central bank have tight control over the money supply? What are the factors that weaken the link between the central bank policies and changes in the money supply?
10. Would a high (in the limit, 100 percent) reserve requirement imposed on banks strengthen central bank control over the money supply? If so, why do central banks never impose very high reserve requirements?
11. Discuss what will happen to the monetary base and the money supply if the central bank starts paying interest, say at the Treasury bill rate, on commercial bank deposits with it. If it does so, would the interest elasticity of money supply be higher or lower than it would be under a regime where interest was not paid on such deposits?
12. Specify the behavioral money supply function that captures the behavior of the central bank, the public and the banks in the money supply process. Discuss and compare its likely interest-rate elasticities for M1 and M2.
13. Let *BR* (bank reserves) be:

    $$BR = kD + ER$$

    Assume:

    $$ER = f(R, R_d)D \quad \partial ER/\partial R < 0, \partial ER/\partial R_d > 0$$

    where ER is excess reserves, R is the market interest rate, R_d is the central bank's discount rate and the meanings of the other symbols are as specified in the chapter. Also, let:

    $$C/D = c$$

    Derive the money supply $M^s(= C + D)$ as a function of k, c, R, R_d and M0 (the monetary base), and specify the signs of its partial derivatives.
14. Given the information in the preceding question,

    (i) If $k = 0.05$, $ER = 0.01D$, $c = 0.2$, $C + D = 0.03\ y$, derive M0 as a function of y. Also, derive C, D and BR as functions of y.
    (ii) Suppose financial innovations shift the demand for currency such that c becomes 0.1. Derive the monetary base as a function of y. How could the central bank offset the impact of this shift in currency demand on the money supply? Discuss.

15. Specify a money supply function in terms of the currency and reserve ratios and the monetary base. Let the real money demand function be:

$$m^{\mathrm{d}} = m^{\mathrm{d}}(y, R) = m_y y \; - m_R R + FW_0 \quad \partial m^{\mathrm{d}}/\partial y > 0, \partial m^{\mathrm{d}}/\partial R < 0, \partial m^{\mathrm{d}}/\partial W > 0$$

where FW_0 is financial wealth.

(i) Given the need for money market equilibrium, show how the central bank can use M0 as its monetary policy instrument for changing $Y(= Py)$.
(ii) Given the need for money market equilibrium, show how the central bank can use (a) R, (b) R_d, as its policy instrument for changing Y.
(iii) Given the need for money market equilibrium, how would an increase in the prices of stocks, which increases FW, change M0, R and R_d?

16. How would you formulate your money supply function for estimation in an empirical study? What would be your definition of the money supply variable? What would be the arguments of your function? Justify each one. Comment on the a priori relationships that you would expect between your independent variables and the money supply. Compare your supply function with some others estimated in the literature, and comment on the differences.
17. For a selected country and using quarterly data, specify and estimate the money supply function. Among your independent variables, include at least two different interest rates, one being the discount/Bank rate and the other a market-determined short-term rate. Check and correct for shifts in the money supply function during the period of your study. Carry out your estimations using the following techniques:

(a) least-squares estimation, with a first-order PAM;
(b) cointegration with an error-correction model.

Discuss your choice of the functional form of the money supply function and your choice of the variables and econometric techniques used, as well as the data and econometric problems you encountered.

Discuss your results, their plausibility and consistency with the theory, and their robustness. Can you explain the estimated shifts in your function by reference to changes in the policies of the central bank?

18. What are the reasons for requiring the use of cointegration techniques in estimating the money supply? What would be the disadvantages of using ordinary least squares for such estimation? If you use both and obtain different estimates, which would you rely on and why?
19. Are there, in general, any problems with the application of cointegration techniques to money supply functions, or with the interpretation of estimates? Can these estimates be relied upon? Discuss.

References

Baghestani, H., and Mott, T. "A cointegration analysis of the U.S. money supply process." *Journal of Macroeconomics*, 19, 1997, pp. 269–83.

Cagan, P. *Determinants and Effects of Changes in the Stock of Money, 1875–1960*. New York: Columbia University Press, 1965.

De Leeuw, F. "A model of financial behaviour." In J.S. Duesenberry *et al.*, eds, *The Brookings Quarterly Econometric Model of the United States*. Chicago: Rand McNally, 1965, Ch. 13.

Friedman, M., and Schwartz, A.J. *A Monetary History of the United States, 1867–1960.* Princeton, NJ: National Bureau of Economic Research, 1963.

Frost, P.A. "Banks' demand for excess reserves." *Journal of Political Economy*, 79, 1971, pp. 805–25.

Goldfeld, S.M. *Commercial Bank Behaviour and Economic Activity.* Amsterdam: North-Holland, 1966.

Poole, W. "Optimal choice of monetary policy instruments in a simple stochastic macro model." *Quarterly Journal of Economics*, 84, 1970, pp. 197–216.

Rasche, R.H. "A review of empirical studies of the money supply mechanism." *The Federal Reserve Bank of St Louis Review*, 54, 1972, pp. 11–19.

11 The central bank

Goals, targets and instruments

This chapter focuses on the institutional and historical aspects of the goals, instruments and targets of monetary policy as they have been pursued by the central banks of the United States, Britain and Canada – with some material on the newly created European System of Central Banks. This material is intended to widen the discussion beyond the particularities of any one country and to provide some indication of the similarities and varieties of central bank practices in the pursuit of monetary policy.

Key concepts introduced in this chapter

- Central banks' mandates
- The potential multiplicity of goals of central banks
- Open market operations
- Required reserves
- Discount/bank rate
- Credit controls
- Moral suasion
- Selective controls
- Overnight loan interest rate
- Federal Funds rate
- Administered interest rates
- Currency boards
- The competitive supply of money

Economic theory has long recognized the impact of monetary policy upon most of the important macroeconomic variables such as output, employment, growth and prices.[1] Therefore, in most countries, the control of the money supply and the manipulation of interest rates, in so far as these are possible, is entrusted to the central bank rather than left to market forces.[2] This chapter looks at the basic practical and institutional aspects of the goals and targets of central bank policies, and related issues such as the regulation of

1 See Chapter 2 for the economic heritage on this, and Chapters 13 to 16 for a modern treatment.

2 However, there are economists, belonging to the new monetary school, who espouse competitive issue of money by private institutions. Such issue of money, even bank notes, by private firms was common in most countries until the twentieth century.

financial intermediaries. Institutional arrangements and practices are specified for the Federal Reserve System of the United States, the Bank of Canada, the monetary arrangements in Britain and the European System of Central Banks, which is the federated central bank for the European Union. The intention in presenting this material on several countries is to show the common elements as well as the diversity of monetary arrangements among a group of countries.

Section 11.1 examines the historically multiple goals of central banks and Section 11.2 investigates their evolution to the present goals, which are price and stability, with output at full employment. Section 11.3 reviews the instruments by which central banks conduct monetary policy. Sections 11.4 and 11.5 focus on the issues of competition and regulation of the financial sector and interest rates. Section 11.6 provides information on the monetary conditions index, which is used as a guide to monetary policy. Section 11.7 relates the evolution of goals of the central bank to the Taylor rule. Section 11.8 appends a brief discussion of currency boards, which represent an alternative institutional arrangement to central banks.

11.1 Historic goals of central banks

The central banks of different countries usually have their own distinctive and somewhat different sets of goals in their mandates from their respective legislative authorities. However, as we shall see in this section, there is also a high degree of similarity in the goals, broadly defined, among them. Further, the mandate assigned to a given central bank is normally broad enough to allow it a great deal of latitude in the goals it does choose to pursue in practice. We illustrate the types of goals usually assigned to central banks by looking at those for the USA, Canada and the UK.

Original mandate of the Federal Reserve System in the USA

The Federal Reserve System, referred to as the Fed, is the central bank of the United States. It was set up in 1913 and has a Board of Governors[3] with a Chairman at its head. Its monetary policy is set by the Federal Open Market Committee (FOMC), which consists of the Board of Governors and five of the presidents of the twelve Federal Reserve Banks. FOMC sets the Federal Funds rate, which is the rate at which commercial banks trade reserves with each other through overnight loans.

A publication of the Federal Reserve System of the USA listed the broad objectives of the system as:

> To help counteract inflationary and deflationary movements, and to share in creating conditions favorable to sustained high employment, stable values, growth of the country, and a rising level of consumption.

It might have also added the additional objective of promoting a favorable balance of payments position. The list of economic goals in the mandates assigned until the 1980s to most central banks was very similar to the above multiplicity of goals.

3 The seven governors are appointed for 14-year terms, which contributes to their independence from the government of the day. The Federal Reserve System is made up of 12 regional Federal Reserve Banks, located in different parts of the USA.

Original mandate of the Bank of Canada

The Bank of Canada was set up in 1934. It has a board of directors and a Governor, who decide on monetary policy. The preamble to the Bank Act of 1934, setting up the Bank of Canada, stated that the mandate of the Bank was to be:

> To regulate credit and currency in the best interests of the economic life of the nation, to control and protect the external value of the national monetary unit and to mitigate by its influence fluctuations in the general level of production, trade, prices and employment, so far as may be possible within the scope of monetary action, and generally to promote the economic and financial welfare of Canada.

This preamble required the Bank of Canada to use monetary policy to achieve multiple goals. There was an implicit assumption behind the preamble that it was within the Bank's power to affect not only the rate of inflation and the exchange rate, but also the real – and not merely the nominal – variables of output and employment. On the latter, the assumption was that the Bank could affect the short-run values of these real variables and thereby influence their fluctuations.

Evolution of the Bank of England and the goals of monetary policy in the UK

The Bank of England was founded as a private commercial bank in 1694.[4] Although a privately owned bank until 1946, it acted from the very beginning as the banker for the British government, under a business arrangement entered into in exchange for large loans made by the Bank to the British government through its purchase of government bonds. It was given the monopoly of (future) note issue in 1844, when it also withdrew from commercial banking. Its notes were made legal tender and convertible into gold at a fixed exchange rate. It evolved into a central bank through custom and practice in the eighteenth and nineteenth centuries, only gradually increasing its responsibility for maintaining orderly conditions in the money markets and influencing the policies and practices of the other commercial banks.

Given its origin as a private bank and its gradual evolution in practice into a central bank, there was no explicit legislated mandate for the Bank of England to pursue monetary policy in order to achieve specific national macroeconomic goals, even though it interacted closely with the government. Its primary goals through the eighteenth and nineteenth centuries seemed to be mainly related to maximizing its own profits and preserving its own solvency. This was consistent with the tenor of traditional classical ideas, which did not possess a theory of monetary policy for manipulating the economy and did not espouse an active monetary policy for stabilizing it.

Since the nationalization of the Bank of England in 1946, the relationship between the Bank of England and the government – represented by the Chancellor of the Exchequer, who is the government minister in charge of the Treasury – has gone through two distinct phases. From 1946 to 1997 the government had statutory power over not only the goals but also the use of instruments of monetary policy, though the day-to-day operations and the normal business were left to the Bank. The Bank made recommendations to the Chancellor, who had the final decision on the goals being pursued. The Bank implemented the policies defined by

4 Its ownership was nationalized in 1946.

the Chancellor, though with some discretion over the timing of implementation of decisions. Consequently, the goals pursued for the economy through monetary policy were ultimately those of the government and depended upon the preferences of the party in power. In 1997, however, the Bank was given operational independence in implementing monetary policy, but the power to set the ultimate goals of monetary policy was retained by the Chancellor and, therefore, by the government.

Given this division of powers in Britain between the Bank of England and the government, it is appropriate to use the term "monetary authority" to encompass both of them in their joint roles of setting the goals and pursuing the implementation of monetary policy. By comparison, in the United States and Canada, the monetary authority will simply be their respective central banks.

In practice, the historic goals of the monetary authority in the UK were very similar to those of the central banks in the USA and Canada. From 1946 to the early 1980s, these policies were based on a wide set of goals, including lower unemployment, higher growth, lower inflation and the maintenance of the exchange rate, under the notion that it was possible to achieve several goals or at least trade off among them through monetary policy.

Mandate of the European Central Bank

The gradual bonding of the European countries during the postwar years and their eventual merger into the European Union in the 1990s resulted in their monetary unification in that decade. The central element of this unification is the European System of Central Banks (ESCB), established under the Maastricht Treaty of 1992. The ESCB consists of the European Central Bank (ECB), based in Frankfurt, and the national central banks of the member countries, and is federalist in structure. The central decision-making body on monetary policy is the ECB's Governing Council, composed of the national central bank governors[5] and the Executive Board[6] of the ECB. The Executive Board, besides running the day-to-day operations of the ECB, carries out the decisions of the Governing Council and coordinates their implementation by the national central banks in their respective countries.

Since the European Central Bank was set up recently, its mandate reflects current thinking on monetary policy. Its charter states that the "primary objective" of the European System of Central Banks (ESCB) should be to "maintain price stability." The charter also establishes the complete independence of the ECB and the national banks from their various governments, so that they are protected from having to take instructions from governments.

The ESCB uses changes in interest rates as the main operational tool of monetary policy, but continues to attach importance to the evolution of monetary aggregates, especially M3, as a guide for its policies.

Additional mandates of the central banks in the LDCs

Most less-developed countries are unable to raise the revenue internally to cover their fiscal, with the result that they continually incur large fiscal deficits. Some of these tend to be

5 Each of them is appointed for a minimum term of five years and cannot be dismissed by their respective governments.

6 The Executive Board has a President and five members, appointed by the European Council for eight year, non-renewable, terms.

covered by foreign borrowing by the government, but very often there is still a remainder that needs to be financed. While such deficits in the richer and financially developed nations are normally covered by the government borrowing directly from the public through the issue of short- and long-term bonds, the financial markets in the LDCs are too thin to support much governmental borrowing. Given this constraint, many governments in the LDCs rely upon the central bank to directly finance the remaining deficit through increases in the monetary base, or to do so indirectly by the sale of government bonds to the central bank.

Hence, while the LDCs have broadly similar mandates for their central banks as in the industrialized countries, an additional one in practice is that of financing the fiscal deficit. The justification sometimes given for this is that of national interest. In some cases, the deficit is related to massive development programs, so that the central bank's financing of the deficit is further claimed to be a contribution to national economic development. However, such a practice makes monetary policy subservient to fiscal policy, and has implications for the independence of the central bank from the government and for its control of inflation.

11.2 Evolution of the goals of central banks

Revision in economic theory and the implied limitation on goals

The 1970s were a period of expansionary monetary policy but were accompanied by stagflation in Western economies. This combination produced a period of increasing doubts about the relevance and validity of Keynesian policy prescriptions, which, in turn, proved to be a fertile ground for the emergence and acceptance of resurgent neoclassical theories. An early element in this resurgence was in the revision of the Phillips curve to the expectations-augmented Phillips curve (see Chapter 14), often associated with Milton Friedman. Friedman (1977) argued that there was a short-run trade-off only between the unemployment rate and the deviation of the inflation rate from its expected level. But there was no long-run trade-off between unemployment and inflation. Hence, monetary policy had a very limited ability to change the unemployment rate. The 1970s also saw the introduction of the hypothesis of rational expectations. Lucas, Barro, Sargent and Wallace, Kydland and Prescott and others in the 1970s and 1980s laid the foundations of the modern classical model with rational expectations and the neutrality of systematic monetary policy as its core elements (see Chapters 14 and 16). Only random monetary policy could affect short-run output but such a policy would be meaningless. Systematic increases in money supply, under conditions of symmetric information between the monetary authority and the public, would be anticipated but could not bring about the deviation of employment from its equilibrium level.

In the 1980s, the impact of this theoretical revision in the scope of monetary policy was to persuade the central banks of many countries to abandon the multiplicity of goals in favor of a heavy and sometimes sole focus on controlling the rate of inflation. While formal legislative revision of the traditional mandates of the central banks was rare, in practice there was a considerable reduction in the emphasis on using monetary policy to change unemployment and output. The primary goal of monetary policy became a low inflation rate and the resulting policy became one of "inflation targeting." The success of the inflation targeting policy since the 1980s in reducing inflation on a long-term basis, accompanied by increasing output and employment, has led to a general adoption of inflation targeting as the primary objective

of central banks in many countries. As of 2002, inflation targeting had been adopted in 22 countries, including Canada and New Zealand (which had started this policy), the UK and the USA. However, the objective of ensuring full employment has not disappeared, since both inflation targeting and output targeting are components of the popular Taylor rule (see Chapters 12 and 13) for setting interest rates.

Evolution of the goals of the monetary authorities in Britain

As discussed earlier, the British monetary authority from 1946 to the early 1980s possessed a multiplicity of goals. While the goal of price stability had been one of those goals and had received increasing emphasis by the end of the 1980s, the sole focus on it was made explicit in 1992 when the Chancellor of the Exchequer announced the adoption of explicit inflation targets (1 percent to 4 percent) aimed at achieving long-term price stability. This adoption of an explicit target for the inflation rate represented an explicit abandonment of other goals such as those on unemployment and output growth, exchange rate stability and business cycle stabilization.

The Chancellor and the Bank periodically set the inflation target. The official inflation target was changed in 1995 by the Chancellor, with the agreement of the Bank of England, to a point target of 2.5 percent. It is currently (2008) at 2 percent, along with a set range for acceptable fluctuations in inflation. The primary objective of monetary policy in Britain is now explicitly that of price stability and only secondarily growth and employment objectives. The implementation of monetary policy, such as the setting of the Bank rate, since 1997 has been left to the Bank and its Monetary Policy Committee. Goodhart (1989, 1995) provides an excellent statement of central banking from a British perspective.

Evolution of the goals of the Bank of Canada

In the late 1980s, the Governor of the Bank of Canada publicly argued that the mandate of the Bank should be changed to focus only on price stability as its mandated goal. The proposal was considered in 1992 by a parliamentary committee, which decided to leave the Bank's mandate as it had been specified in the Bank of Canada Act of 1934, i.e. with a multiplicity of goals. However, several successive Governors of the Bank in the late 1980s and the 1990s have advocated and in practice consistently focused solely or mainly on the goal of price stability or a low inflation rate. Since 1991 the Bank has announced explicit targets, with an average of 2 percent and a range of 1 percent to 3 percent, for the core inflation rate. These targets have been set jointly by the Bank of Canada and the Government of Canada.[7] Movements in other variables such as the exchange rate and asset prices are taken into account to the extent that they affect the rate of current or future inflation. In the case of a deviation of the actual inflation from the 2 percent target, the Bank normally aims at bringing inflation back to its target over a six to eight-quarter period. The Bank uses a Monetary Conditions Index (MCI) (explained later in this chapter), which is a weighted sum of the interest rates and the exchange rate, as an operational guide. It uses the overnight loan rate for reserves as its operational target, with a range of 50 basis points. The Bank Rate – on its loans to

7 In Canada, the inflation target is set every five years. In 2006, the government and the Bank of Canada renewed the inflation target of 2 percent, with a range of 1 percent to 3 percent.

banks and some other financial institutions – is now set at the upper limit of the target range.

New Zealand's experiment on the goal of price stability

New Zealand was the first country to explicitly adopt inflation targeting. After an extended period of double-digit inflation for much of the late 1970s and 1980s, as well as unsatisfactory growth, major legislative changes were made to the country's monetary arrangements in the mid-1980s and in 1990. Among these was the grant of a limited degree of independence to the central bank, the Reserve Bank of New Zealand, to formulate and implement monetary policy so as to maintain price stability. However, on the formulation of the goals of monetary policy, it required the Minister of Finance and the Governor of the central bank to jointly establish the specific inflation target, its range and the inflation index to be used for the target. This information is communicated to the public. These agreements are renegotiated at certain intervals and leave a measure of flexibility in meeting changing economic conditions. The target ranges for inflation were 3–5 percent from 1990 to 1992, 0–2 percent from 1992 to 1996 and 0–3 percent since 1996. Permissible breaches of the established target are allowed in special circumstances, such as natural disasters, changes in indirect taxes and significant relative price shocks. The Reserve Bank of New Zealand currently uses the *Monetary Conditions Index*, a concept pioneered by the Bank of Canada. This index is a weighted sum of an interest rate measure and the exchange rate, and is used as a guide for monetary policy.

The New Zealand pattern is similar to that in Britain, and to some extent that in Canada. Its central bank has operational independence but not total independence over the ultimate goals of monetary policy. The goal is limited to that of price stability by legislation, and the legislation requires that the target range be set jointly with the government.[8]

Recent pattern of the goals of monetary policy in the United States

The pursuit of goals by the Federal Reserve System changed in the 1980s and 1990s in a manner similar to that in the other countries discussed above, from the pursuit of multiple goals during the pre-1980 period to that of price stability. One difference between the Fed and the British and Canadian monetary authorities is that the Fed does not set explicit targets for the rate of inflation, though its pursuit of a low rate of inflation consistent with price stability is not in doubt and is often asserted by the chairman of the Board of Governors of the Fed. Taylor (1993) argued that the Fed had effectively moved to a monetary policy rule (see the Taylor rule later in this chapter and in Chapters 12 and 13) incorporating inflation targeting.

The Fed is more genuinely independent of the United States' President and government than are the Bank of England or the Bank of New Zealand of their governments, in terms of both the formulation of its goals and their pursuit through its monetary policies. There is no question of the government issuing formal instructions to the Fed on the further pursuit of monetary goals or instruments, of requiring it to write open letters to the government explaining its actions, or of penalties being imposed for a failure to achieve price stability.

8 What is distinctive about the New Zealand case is that the Governor of the Bank can be held responsible for failure to achieve the announced targets, and sanctions may be imposed for such failure.

11.3 Instruments of monetary policy

The central bank pursues its monetary policy through the use of one or more of the instruments at its disposal. The mix of the instruments used depends upon the structure of the economy, especially the financial system, and tends to depend on the stage of development of the bond and stock markets. The most common instrument among developed countries is often the change in interest rates, which is usually supported by open market operations.

11.3.1 Open market operations

Open market operations[9] are the purchase (or sale) by the central bank of securities in financial markets, resulting in corresponding increases (decreases) in the monetary base.[10] Countries with well-developed financial markets and extensive amounts of public debt traded in the financial markets usually rely on such operations, which, along with the shifting of government deposits between the central bank and the commercial banks, are the most important tool for changing the money supply. However, this does not hold a corresponding position among the countries around the world since its prominent position requires certain preconditions to be met. The most important of these are:

1 The financial structure of the economy should be well developed, with most of the borrowing and lending being done in the organized financial markets of the country itself.
2 There should be a relatively large amount of the securities of the kind that the monetary authorities are willing to purchase. These are often, though not always, government securities. A large public debt is thus generally essential.
3 The financial system and markets of the country should be broadly independent of those of other countries. Very open economies with perfect capital flows and fixed exchange rates do not possess such independence. To take an extreme example, different states or regions within a country do not possess an independent financial system and cannot pursue a monetary policy independent of the country. Similarly, members of a currency or monetary bloc with fixed exchange rates among the member countries cannot pursue monetary policies independently of that of the bloc. An example of such a bloc is the European Union with fixed exchange rates among the national currencies of the member countries.

Financially underdeveloped economies generally fail to satisfy the first condition. Individual countries in an economic and/or political union, such as the European Union, may not fully meet the third condition, so that their national central banks cannot independently pursue such operations, but may do so in support of the monetary policy set by the central

9 The preponderant part of such operations on the monetary base is usually of a defensive nature, aimed at smoothing undesired variations in the monetary base.

10 Often, as in the USA, the central bank may conduct open market operations only a few times each year to meet the economy's longer term liquidity needs, while leaving borrowing by commercial banks from the central bank to adjust their day-to-day needs. Different countries have different types of arrangements for such temporary borrowing. In the USA, such borrowing is often done through repurchase agreements between the central bank and the borrowing bank.

bank of the overall union.[11] Many countries of the world fall into one or more of these categories. Even in the United States, open market operations were rarely used prior to the Second World War. Most countries pursue other tools of monetary policy, to supplement, or as an alternative to, open market operations.

Shifting government deposits between the central bank and the commercial banks

The central bank almost always acts as the government's bank, keeping and managing the government's deposits. Increases in these deposits with the central bank through payments by the public to the government out of their deposits in their commercial banks reduce the monetary base, whereas decreases in such deposits because of increased payments by the government to the public increase the monetary base. One way of avoiding changes in the monetary base because of payments to the government or receipts from it, is for the government to hold accounts with the commercial banks and use them for its transactions with the public. The resulting increases and decreases in government deposits with the commercial banks do not change the monetary base, but transfers of these deposits to the central bank reduce this base.

In Canada, the Bank of Canada manages the distribution of government deposits between itself and the chartered banks in Canada as a way of manipulating the monetary base and therefore as a tool of monetary policy akin to open market operations. In current practice, such shifting of balances is more convenient and has become more important than open-market operations for changing the monetary base over short periods.

11.3.2 Reserve requirements

The imposition of reserve requirements[12] has historically been a common tool of controlling monetary aggregates for a given monetary base. In cases where the markets are too thin for viable open-market operations, or the monetary base cannot be controlled for some reason, the monetary authorities often attempt to limit the creation of reserves by the banking system through the imposition of, or changes in, reserve ratios against demand deposits and sometimes also against other types of deposits. These ratios can range from 0 percent to 100 percent, though they are often in the range of 0 to 20 percent, with changes in the required ratio being often of the order of 0.25 percent or 0.5 percent.

Until 1980, the USA had a complex system of reserve requirements for its banks, with different requirements for banks that were members of the Federal Reserve System and those that were not, between banks in large cities and others, etc. In 1980, the US Congress imposed much greater uniformity on the depository institutions, including banks, thrift institutions and credit unions. The Fed was given the power to set the reserve requirement between 8 percent and 14 percent on transactions deposits and raise it to 18 percent in special cases. The reserve requirements on personal time and savings deposits were eliminated. In 1998, the reserve requirement was 10 percent on checkable deposits if the bank had checkable deposits above a certain amount, and 3 percent if these deposits were below this limit. There was no positive reserve requirement on non-checkable time deposits.

11 In the past, this was also often true of the colonies of imperial countries.

12 In most countries, the required reserves have to be held by the bank in currency or in deposits with the central bank. These deposits normally do not pay interest.

Canada had reserve requirements of 5 percent against demand deposits in chartered banks from 1935 to 1954, though the banks often kept much higher reserves (sometimes over 10 percent). From 1954 to 1967 the reserve requirement was 8 percent, with the Bank of Canada having the power to raise it to 12 percent, though this power was never used. By the Bank Act of 1967 the required ratio was raised to 12 percent against demand deposits, with 4 percent on notice deposits in Canadian dollars, but the power of the Bank to vary them was eliminated.[13] In 1980 the required reserve ratio against demand deposits was fixed at 10 percent, with lower ratios against other types of deposits. In early 1992 Canada, in an environment of a highly stable and well developed financial system, abolished reserve requirements on its banks, leaving them to determine whatever amounts of reserves they wished to hold. However, they still have to maintain non-negative settlement balances with the Bank on a daily basis, with any negative balances being offset by overdrafts from the Bank at the bank rate. The average reserves held by the Canadian commercial banks are now usually less than 1 percent of demand deposits and sometimes fall as low as 0.1 percent or even lower.

In Britain, after 1945, the London clearing banks, which are the main clearing banks in Britain, adopted the practice of keeping a minimum reserve ratio of 8 percent of their deposits. Changes in it were never required as an instrument of monetary policy. After 1971, the banks agreed to keep an average of 1.5 percent of their eligible liabilities (mainly their sterling deposits) in non-interest bearing deposits with the Bank of England. Even this requirement was eliminated in 1981, so that the reserve requirement in Britain became zero percent. In 1999, the average ratio of the banks' non-interest bearing deposits with the Bank of England to the sterling deposits placed with them was about 0.15 percent.

The difference between the reserve requirements in Britain and Canada versus those in the USA is illustrative. Part of the reason is historical patterns. But part is also due to the nature of the British and Canadian banks, which tend to be large and country-wide with few failures in the past. Both countries display sufficient confidence in the solvency of their banks to eliminate positive reserve requirements. Although some of the US banks are among the largest in the world, many, if not most, of the US banks are relatively small, confined to a state or a region, with a pattern of bank failures among such banks. Higher reserve requirements contribute to their solvency and the public's confidence in them.

Table 11.1 shows the reserve requirements in 1998 in the G-7 countries, along with the length of the averaging period.

As Table 11.1 shows, there are considerable variations in the reserve requirements among this group of countries. However, the countries with large oligopolistic banking systems tend to have very low, almost zero, reserve requirements. All countries allow some averaging period for meeting the reserve requirements, though it is only one day in the British case. Without averaging, or with averaging over only one day, the daily shocks to the banks' deposits and consequently their demand for liquidity result in sharp movements in the overnight interest rates, unless the central bank can manage to monitor these daily and to

13 From 1967 to 1992, the banks also had to meet a secondary reserve requirement. The reserves for this purpose were defined as the excess of primary reserve requirements plus Treasury bills plus loans to investment dealers, to be held against Canadian dollar deposits plus foreign currency deposits of residents with banks in Canada. The secondary reserve requirement was abolished in June 1992.

Table 11.1 Reserve requirements, 1998

	UK	*USA*	*Germany*	*France*	*Italy*	*Japan*	*Canada*
Required ratio	zero	3–10%	1.5–2%	0.5–1%	15%	0.05–1.3%	zero
Length of averaging period	1 day	2 weeks	1 month	1 month	1 month	1 month	4–5 weeks

effectively offset the results of such shocks. Since it is not possible to do so on a daily basis in many countries, averaging over several weeks is the general pattern.

In the Western economies, changes in the required reserve ratios are now either no longer available as an instrument of monetary policy or not used in practice for this purpose.

11.3.3 Discount/bank rate

In most countries, the monetary authority – normally the central bank – has the power to determine, directly or indirectly, the interest rates in the economy.[14] Critical interest rates can be set directly by fiat, determined through instructions issued to the commercial banks, or influenced indirectly by the central bank varying the rate at which it lends to the commercial banks. In the more usual case in market-oriented economies, the market rates are influenced through the *discount rate* at which the central bank lends to the banks and other designated financial intermediaries and by the market overnight loan rate for reserves. Canada, the UK and the USA have traditionally followed this method.

The use of interest rates as the major operating instrument of monetary policy occurs because interest rates play a pivotal intermediate role by which investment and therefore aggregate demand in the economy can be influenced. Further, it is argued by some economists that the economy has numerous substitutes for M1 and M2,[15] so that controlling these aggregates through open-market operations or the reserve requirements of commercial banks only leads to substitution away from them, without necessarily a significant impact on investment and aggregate demand.[16] Further, in recent years, because of numerous financial innovations, the demand functions for money have proved to be unstable, so that many central banks prefer to target interest rates, and influence them through their discount rate, rather than target monetary aggregates as the main operational tool of monetary policy. Chapters 10 and 13 provide the justification for this policy for controlling aggregate demand if the shocks are mainly from the monetary sector rather than from the commodity one.

14 This power to determine the domestic interest rates varies between closed and open economies and can be very limited for small open economies. The use of such a power is also more common in LDCs than in the developed economies.

15 This claim is often associated with the Radcliffe Report in the United Kingdom in the late 1950s. It claimed that the economy was awash in liquidity, which included trade credit and short- and medium-term bonds, and money was only a small component of it. Restrictions in its supply merely led to its substitution by other liquid assets. This extreme version of the argument was abandoned since the money-demand functions proved to be quite stable for the 1960s and 1970s, but it can still be used in a milder form.

16 However, Milton Friedman and other monetarists argued in the 1960s that the demand function for money was stable, and more stable than the investment multiplier, so that controlling the money supply rather than the interest rates (and through them, investment) was preferable. The experiment in using monetary aggregates as targets was briefly tried during the late 1980s but was discontinued because of the instability of the money-demand function.

The central bank, in setting or changing its discount rate, indicates its willingness to let the commercial banks determine the extent of borrowing from it and thereby changing the monetary base in the economy. Its target for the overnight loan rate for reserves determines the rate at which commercial banks borrow from each other. Any announced changes in these rates act as indicators of the bank's future intentions about the economy's interest rates that it will support through its open-market operations, and therefore serve as an indicator of the future stance of monetary policy. The commercial banks and other financial intermediaries usually, though not always, follow the lead given by the discount and the overnight loan rate changes to alter their own interest rates[17] – such as the prime rate, the personal loan rates and the mortgage rates – as well as in their purchases and sales of market instruments. This behavioral pattern results in a shift of the interest rates throughout the economy, while leaving the spread between any pair of rates to market forces. Conversely, the central bank's refusal to change these rates in the face of rising market rates serves to dampen the latter.

This discount rate is called the bank rate in Canada and the UK. In the UK, as explained earlier in this chapter, since 1997 the Bank of England, through the Monetary Policy Committee, has had the operational independence to set this rate. Until 1971, a sort of cartel arrangement among the British commercial banks linked the market interest rates on various types of bank deposits to the bank rate. The abolition of this cartel in 1971 made the market rates more responsive to market forces, though the bank rate set by the Bank of England still continues to be the core rate for the financial markets and changes in it are the major operational instrument of monetary policy. Since the British banks as a whole need to achieve balance at the end of each day, the Bank of England can choose on a daily basis the interest rate at which it will provide additional funds to the banking system. Changes in this rate prompt the banks to change the base rates at which they lend to their customers, so that the changes in the Bank's own lending rate cascade through the various interest rates in Britain.

In Canada, the bank rate (at which the Bank of Canada lends to commercial banks) was set by the Bank until 1980. In the 1980s and the early 1990s, it was fixed at 1/4 percent above the 91-day Treasury bill rate at its weekly auction of government bonds. Since it was above the Treasury bill rate, it was considered to be a "penalty" rate in the sense that it imposed a net loss on the borrowing bank, which had the option of obtaining the needed funds more cheaply by selling from its holdings of Treasury bills. The Bank influenced the Treasury bill rate through its own bids for Treasury bills. Since 1994, the Bank has been setting the overnight loan rate – that is, the rate on trades in reserves – with a band of 50 basis points, as an operational target. Since 1996, the bank rate has been set at the upper limit of the operating band specified for the overnight loan rate. Setting the bank rate at this upper limit makes borrowing from the Bank more expensive than in the commercial market for reserves and is meant to persuade commercial banks to meet their reserve needs through their borrowing of reserves in the private markets. However, borrowing from the Bank is treated as a right of the banks rather than a privilege. In any case, the Canadian banks consider borrowing from the Bank of Canada as sending out a signal that they have liquidity problems and are reluctant

17 In Canada, the UK and the USA, the ability of the central bank to support its interest rate policy is credible, so that the market rates adjust immediately to reflect the change in its rate. However, this is not necessarily the case in many other countries, especially in those with underdeveloped financial markets.

to resort to such borrowing. Any advances are normally for only a few days and often overnight.

In the United States, the discount rate is frequently below the market short interest rates. Since banks can make a profit from borrowing from the Fed and then buying market instruments, keeping the discount rate below the market rates becomes an incentive for commercial banks to borrow from the Fed. However, the Fed treats borrowing from it as a privilege rather than a right. Frequent borrowing from the Fed can lead to its refusal to lend again and would invite closer scrutiny of the borrowing bank's accounts and policies. Further, a bank may view its borrowing as a signal to the public that it is in dire financial need and is not able to conduct its affairs properly, so that it may be reluctant to borrow.[18]

A change in the discount rate can serve as an instrument of monetary policy in three respects:

1 It affects the amount of borrowing – which changes the monetary base and the money supply – from the central bank.
2 Changes in it – or lack of a change when one was expected – act as a signal to the private sector of the central bank's intentions about monetary policy.
3 The central bank's control over its discount rate provides it with considerable control over the interest rates in the economy.

The latter two are now the relatively more important reasons for the use of the discount rate as an instrument of monetary policy.

Central bank as the lender of last resort

Borrowing from the central bank at the discount rate is associated with the notion of the central bank acting as the *lender of last resort* in the economy. While commercial banks with inadequate reserves can borrow from those with surpluses, a reserve shortage in the financial system as a whole cannot be met in this manner and could force the economy into a liquidity and credit crunch. The *discount window* – i.e. the ability to borrow from the central bank – is therefore a "safety" valve for the economy.

The discount window also acts as a safety valve for an individual bank that needs reserves but is unable or unwilling to borrow from private financial institutions. However, in the United States, borrowing from the central bank invites the scrutiny of the central bank into the borrowing bank's management of its affairs and acts as a disincentive to frequent borrowing from the central bank, as against borrowing in the market. Further, banks are not permitted to make chronic use of the discount window for meeting liquidity needs.[19]

18 Given such an attitude, at the time of the subprime financial crisis in August–September 2007 the Fed, which wanted to ensure the liquidity and solvency of the banking institutions, actively encouraged commercial banks to borrow from it.

19 In Britain, *discount houses* usually act as intermediaries between the banks and the Bank of England. Banks that need funds can draw down their balances at the discount houses, sell securities to them or borrow from them. If the discount houses need funds, they can either borrow or sell securities to the Bank of England at the bank rate (also called the dealer rate). The latter method is now the more common one. The explanation for this indirect method of banks' borrowing from the Bank of England is that direct borrowing by a bank might be seen as a sign of liquidity problems and could reduce the public's confidence in the bank.

Discount rate and interest rate differentials in the economy

The central bank's power to set its discount or bank rate does not extend over the differentials or spreads among the various interest rates in the economy. In particular, the spreads between the commercial banks' deposit rates and the short-term market rates, such as on Treasury bills and money-market mutual funds, are still outside the direct influence of the central bank and depend upon market forces.

From the perspective of monetary theory, the determinants of the demand for M1 and other monetary aggregates are critical for the impact of monetary policy. These demands would depend on both the levels and the differentials among interest rates. Since the latter are mainly outside the influence of the central bank, the impact of the changes in the discount rate on the demand for monetary aggregates is correspondingly reduced.

11.3.4 Moral suasion

Moral "suasion" – rather out-of-date term for "persuasion" – refers to the use of the influence of the central bank upon commercial banks to follow its suggestions and recommendations, such as in exercising credit restraint or diverting loans to specified sectors of the economy. Such suggestions do not possess the force of law, though the threat of converting suggestions into legal orders, if necessary, often backs such suggestions. Moral suasion generally works well in countries with a very small number of large banks and with a long tradition of respect for the judgment and extra-legal authority of the central bank. Although both the UK and Canada are good examples of this, the Bank of England is especially known for its extensive use of moral suasion.

However, moral suasion is not generally appropriate for the large and diverse banking system of the United States, though it has been tried sometimes. An example of the latter occurred in 1965 when the President and the Federal Reserve laid down guidelines to limit foreign borrowing. This was fairly well adhered to by the member banks, but represented a rare usage of this tool in the USA. This term is also sometimes associated with the rules imposed by the Fed on banks that attempt to borrow too frequently from it and is, therefore, associated in the US with the use of the discount window.

11.3.5 Selective controls

Selective controls are those with impact on certain sectors rather than upon the overall economy. A common example of these is credit controls. The usual reason for such controls is that social priorities may differ from private priorities. Thus the government may wish to divert funds to exports, housing, agriculture, state and local governments and to industries believed to be essential to national development. This can include giving special rediscounting privileges to private commercial export bills. Some central banks also favor housing and agriculture through favorable discount provision and direct credit controls, with such support provided under the regulations and guidelines given by the central bank to the commercial banks. However, such support in the United States, Canada and the UK is generally fiscal, in the form of tax exemptions, subsidized loans from the government, etc., rather than being an aspect of monetary policy.

Another reason for selective controls is to curb the destabilizing nature of certain sectors or to use the critical position of certain sectors for stabilization purposes. For example,

on the former, the Federal Reserve limits the stock market credit extended by banks and brokers on the purchase of securities by setting minimum-margin requirements. These specify the minimum down payment at the time of purchase. This requirement was, for example, 70 percent in 1968 for stocks registered on national security exchanges, so that purchasers of such stocks could borrow only up to 30 percent of the purchase price from banks or brokers. The Federal Reserve can raise these requirements up to 100 percent.

Another example of such controls is those on consumer credit. These often specify, for designated durable consumer goods, the minimum down payment at the time of purchase and the length of time over which the balance has to be paid. Such controls are exercised in some countries and often go under the name of installment-credit or hire-purchase controls. In the USA, the Fed was given the power to impose such controls in the Second World War, in the Korean War and briefly in 1948–49, but does not now possess such a power.

11.3.6 Borrowed reserves

Funds borrowed by commercial banks from the central bank are called borrowed reserves. Those acquired through their sale of securities or through the deposits with them by the public are called non-borrowed reserves.

The distinction between borrowed and non-borrowed reserves is important since commercial banks usually are less willing to lend as high a proportion of the former as of the latter. This occurs since the former are borrowed for short periods and carry the discount rate, which is usually higher than the commercial paper rate. In addition, there is often a reluctance to borrow, or to borrow repeatedly, since banks view doing so as creating doubts in the public's mind about the borrowing bank's ability to manage its affairs properly. There is, therefore, often a stigma attached to such borrowing. Repeated borrowing also invites closer oversight, which is considered undesirable, by the central bank and other regulatory authorities of the bank's investment policies.

The central bank can vary the amount borrowed from it by varying the discount rate. A decrease in this rate, relative to market interest rates, makes it more tempting for banks to increase their borrowing.

As against borrowing from the central bank, banks can borrow reserves from each other at the Federal Funds rate in the overnight loan market for reserves; this is called the Federal Funds market in the USA. This market enables banks to lend their excess reserves to other banks that are short of reserves. The central bank controls both its discount rate and the overnight loan rate. The central bank manages the latter by open-market operations in the reserves market; an open-market purchase of bonds by the central bank increases the reserves traded in this market and reduces the rate.

To sum up, individual banks can borrow reserves in the overnight loan market from other banks that have excess reserves, through repurchase agreements with the central bank, or from the central bank at the discount rate.

11.3.7 Regulation and reform of commercial banks

In addition to its control of the economy through monetary policy, a major function of central banking and its related regulatory authorities is to ensure the continued health of the financial system as a whole and of the individual banks. The latter can involve various types

of regulations, such as restrictions on the amount of credit and the payment of interest,[20] restrictions on the type of investments that the financial institutions can make, etc.[21]

11.4 Efficiency and competition in the financial sector: competitive supply of money

11.4.1 Arguments for the competitive supplies of private monies

It is a major contention of economic theory that production and exchange are at their most efficient in perfectly competitive markets. Hence, social welfare is maximized by having perfect competition in all sectors, including the financial one. Even if the actual markets cannot be made fully competitive, restrictions on competition are damaging to efficiency. These tenets are applicable not only to the markets for consumer and investment goods but also to the financial markets. As a corollary, some economists have argued that the financial markets would be most conducive to maximizing the economy's output if they were free from administrative regulations on the products that financial institutions can supply and the prices they charge for these. The products supplied by the financial sector are essentially types of financial intermediation, and involve the holdings of assets and the issue of liabilities by the financial intermediaries. The prices involved are interest rates in the financial markets and service charges, etc. imposed by financial intermediaries. Hence, microeconomic theory implies that the various types of financial institutions should be allowed to compete with each other in the various financial markets, such as for demand deposits, savings and time deposits, mortgages, the purchase and sale of shares, mutual funds, trusts management, pension funds, insurance, etc.

Some economists extend this argument to the proposal that the issue of money should also be left unregulated and that there is no need for a central bank.[22] In fact, since the existence of such a bank with its issue of fiat money represents monopoly power over one aspect of money, its existence and the supply of fiat money reduce social welfare. In line with this argument, it is proposed that private, competitive firms should be allowed to issue coins and notes. It is further argued that the power of commercial banks to create inside money in the form of demand deposits and other types of near-monies should not be limited by any imposition of reserve requirements. Nor should there be regulations on the interest rates charged, on ownership, limitations on the encroachment of banks on trust companies, insurance companies, etc., by limiting the products they can supply. As we have mentioned above, the basis for such proposals is the application of the principle of Pareto optimality of perfect competition to the provision of monies and other financial products.

20 Until the early 1980s, the USA had imposed specific limits on the payment of interest on bank deposits. These were phased out in the early 1980s.

21 In the USA, until the early 1980s, savings and loan associations were restricted to making mortgage loans. Until the late 1990s, commercial banks were not allowed to sell corporate securities, while investment banks could do so but were prohibited from commercial banking. The Fed and the Comptroller of the Currency establish the minimum capital requirements (provided by the banks' shareholders) of the banks' assets. To encourage trust in the security of deposits, the Federal Deposit Insurance Corporation (FDIC) insures bank deposits to a specified limit.

22 Selgin (1988) provides a spirited defense of a banking system free from regulation and dismisses the idea behind central bank control of the money supply that competitive commercial banks, free to issue banknotes, would overissue notes and deposits for private gain.

11.4.2 Arguments for the regulation of the money supply

While economists generally accept the propositions on the promotion of competition among the firms in the financial sector, few economists accept the proposals on the abolition of the central bank, on the elimination of its power to issue fiat money or on the elimination of its power to monitor and regulate the financial system to ensure its continued health, etc. The basic reason for this stand is that the health and stability of the monetary sector is considered to be vital to the prosperity and functioning of the macroeconomy. Further, variations in the supply of money are taken to have a strong impact on the real sectors of the economy.

The commercial banking system is inherently fragile and based on trust by the depositors in the viability of the institutions in which they hold their deposits. A purely competitive and unregulated system is prone to fluctuations in the degree of this trust and therefore susceptible to runs by depositors concerned about the security of their deposits. Two basic reasons for this are the fractional reserve practices of banks and their policy of borrowing short and lending long. Since banks keep reserves, in cash or in deposits with the central bank, equal to only a small fraction of their deposits, they cannot at short notice meet a sudden attempt by depositors to withdraw their deposits in cash. This is further exacerbated by the portfolio policies of banks under which large proportions of their assets are in bonds, mortgages, loans, etc., which are difficult to convert into cash at short notice or can only be cashed with significant losses. Individual banks may also be tempted to engage in high-risk investments, with a consequent possibility of losses, which create a loss of confidence in the bank accompanied by a run to withdraw deposits from it.

Given this inherent fragility of a purely private competitive banking system, several measures are taken to ensure the continuation of a high level of confidence in the banking system. One of these is the insurance of individual deposits, usually to a pre-set limit, by a central deposit insurance agency. Another is having a central bank that attempts to anchor the supply of privately created inside money in the economy through its own issue of fiat money, and also attempts to control variations in the aggregate money supply in the national interest. In addition, the central bank tries to ensure confidence in the financial sector through its regulation and monitoring of financial intermediaries, especially commercial banks since they supply the most liquid financial assets in the economy and are the creators of inside money.

Focusing first on the supply of fiat money by the central bank, one reason for its issue by the central bank is to anchor the privately supplied demand and other types of deposits and therefore the money-supply aggregates in the economy. Another reason concerns the seigniorage – that is, the revenue – emanating from new issues of the monetary base. The central bank is a national institution and its profits are added to the fiscal revenues, so that it seems to be the obvious recipient of such seigniorage. Further, in many low-income economies, seigniorage can be a significant proportion of national revenues and are needed for financing government expenditures.

11.4.3 Regulation of banks in the interests of monetary policy

From the perspective of the central bank, an important aspect of its activities is its regulation of the financial institutions in the economy. Part of this regulation is aimed at the control of the money supply in the national macroeconomic interest. Another part is aimed at maintaining a sound financial system and encouraging, if necessary, its growth to fit the financial needs of the economy. This supervision often takes the form of regulations on the ownership of

such institutions, forms of liabilities issued, the kind of assets held and the auditing of their accounts. Such supervision is only of minor interest from the standpoint of macroeconomics provided it is successful in maintaining a stable and adequate financial system. But it is often a substantial part of the activities of the central bank and its related agencies, and can be critical for the solvency and efficiency of the financial system of the country.

Monetary economics is closely concerned with those regulations of the monetary authorities that affect the liquidity of the economy, especially as reflected in the monetary aggregates. As discussed earlier, among these regulations the central bank often specifies the minimum reserves that commercial banks must maintain against demand deposits. The interest rate at which commercial banks can borrow from the central bank is also set by the central bank rather than based purely on a market mechanism. There may also be other conditions imposed on such borrowing. The maximum interest rates that commercial banks may themselves pay on various kinds of deposits are also, in some countries, set by the central bank. There may be, and often are, other areas of regulation of commercial banks' behavior.

The basic reason for the close regulation of commercial banks lies in the fact that they issue demand deposits that are a major part of the money supply, no matter how it is defined. Most of the regulations governing commercial banks are, in fact, aimed at regulating their creation of demand deposits, with the aim of bringing the total amount of demand deposits and hence the total money supply within the control of the central bank.

Historically, banking in the British tradition arose under a set of customary practices and imposed rules that restricted commercial banks to the issue of demand and savings deposit liabilities and the holding of short-term government bonds as assets. Banks were thereby confined to the highly liquid end of the spectrum of financial assets, leaving other specialized financial institutions to the markets for mortgages, insurance, trusts, pension funds, etc. In addition, there were also restrictions on bank ownership by non-bank corporations, as well as on the ownership of the latter by the banks.

This pattern began to change in the second half of the twentieth century, both in the issue of bank liabilities and in their portfolio of assets.[23] The changes became more pronounced during the 1980s and 1990s, with the financial institutions increasingly being permitted to expand into other than their traditional financial markets, as well as to own or be closely associated with financial institutions in other markets. These changes allowed commercial banks to issue mutual funds, act as investment brokers for the buying and selling of shares, sell insurance and manage pension funds – and conversely to allow firms formerly engaged in these markets to offer banking services. The result by the end of the twentieth century was a breakdown in the USA, Canada and Britain of the barriers between types of financial institutions, mergers and eventually larger sizes of the financial firms, as well as much more aggressive competition in the financial markets.

An aspect of the limitations on banks had been the regulation of the interest rates that banks could pay on the demand and savings deposits placed with them. Often this was an attempt to prevent too aggressive a competition for deposits and to ensure the solvency of banks. An example of this was Regulation Q in the USA during the 1950s and 1960s, under which ceilings were imposed by the Federal Reserve on the interest rates paid by its members on deposits, while many other financial institutions were not subject to such limits. In the interests of promoting competition and removing discriminatory restrictions on banks,

23 In Canada, the chartered banks were first permitted to hold mortgages in the 1950s.

such ceilings were first gradually raised and then eliminated in the 1970s and 1980s. The imposition of ceilings on interest rates paid by banks is now rare in financially developed economies, and such ceilings do not exist in Canada, the UK or the USA. But they do exist in many countries, especially among the LDCs.

11.5 Administered interest rates and economic performance

Manipulating interest rates as guidance to monetary policy or as an aspect of short-term stabilization policies is quite different from setting them over long periods in order to achieve some long-run objectives. Among such objectives is that of attempting to increase the long-run growth rate of the economy.

Interest rates represent the cost of investment, which is the increase in the capital stock of the economy and an essential requirement for the growth of the economy's output capacity. Hence, it can be argued that low rates of interest imply higher investment and therefore higher growth rates for the economy. Many countries, especially LDCs, in the second half of the twentieth century, followed this reasoning to set interest rates in the organized markets of their economies below what would have been determined in unregulated markets. These rates were usually not adjusted to the rate of inflation, so that the interest rates that could be charged often fell below the inflation rate, thereby implying a negative real rate of return on loans.

Interest is not only the cost of funds borrowed for investment, it is also the return on savings lent through the financial markets. Neo-classical theory posits a positive relationship between them, so that lower rates of interest imply lower saving. However, the empirical significance of this dependence of saving on interest rates is in considerable doubt. If saving in practice does not depend on interest rates, while investment does so, it could be argued that keeping the rates of interest low would promote growth of the economy on a net basis.

Interest rates, however, also play the role of allocating funds between the various projects and sectors of the economy. With interest rates below their levels for clearing the markets for loans, administrative mechanisms come into play to allocate the limited funds to the greater demand for them. Among these mechanisms are governmental or central bank regulations on the sectors, projects or firms that are to be given credit, rules of the banks themselves, favoritism of bank managers, etc. Corruption often becomes rife in such a context and becomes a basis for the granting of loans. The end result is the misallocation of funds to projects and firms, in which the most productive uses do not always or adequately get the funds. Such misallocation is detrimental to the growth of the economy. Conversely, leaving the interest rates to be determined in the open and competitive markets for loans promotes the efficient allocation of savings to the variety of investments and thereby increases the growth of the economy. This realization by many LDCs in the 1980s and 1990s led to the "liberalization" of interest rates – a term for lifting ceilings or setting them free to be determined by market forces – in many of them. Such freeing of the interest rates from administrative control is very often part of the broader "liberalization" of the economy, through deregulation and decontrol of exchange rates, imports and exports, production and investment, etc., and has resulted in many cases in increasing the growth rates of those economies.

While all economies have an informal financial sector in which borrowing and lending take place other than through the established and regulated financial intermediaries, such a sector in the LDCs is larger and more significant relative to their formal financial sector. This sector is not only outside the purview of central bank control and policies, its spread

between the deposit rates and the loan rates is usually much larger than in the formal sector. The low deposit rates discourage saving while the high loan rates discourage borrowing for productive investments. Therefore, while the informal sector is vital to the economies of the LDCs, policies which force savers and borrowers to the informal sector through restraints on the formal sector tend to reduce saving and investment in these economies. This implies a recommendation for the competitive and efficient expansion of the formal sector relative to the informal one, though not necessarily through statutory restrictions on the latter.

11.6 Monetary conditions index

In 1992, the Bank of Canada defined a *Monetary Conditions Index* or MCI, which is a weighted average of short-term interest rates and the trade-weighted exchange rate of the Canadian dollar. Roughly, a change in the MCI[24] is specified as:

$$\Delta \text{MCI} = \Delta R + (1/3)\Delta\rho$$

where R is the short-term nominal interest rate, interpreted as the 90-day commercial paper rate, and ρ is the effective exchange rate, interpreted as the exchange rate for the Canadian dollar against the 10 major (G-10) currencies. The reason for the one-third weighting of the exchange rate relative to the interest rate is the Bank of Canada's belief, based on empirical work done therein, that a change in interest rates by 1 percent has three times the impact of a corresponding change in exchange rates on aggregate demand in the Canadian economy. Given that the effect is in the same direction for both R and ρ, opposite changes in these variables offset each other's effects on the economy. Hence, if an increase in the exchange rate occurs and if the Bank considers the resulting increase in the MCI to be undesirable, the Bank responds by inducing a sufficient offsetting reduction in interest rates to keep the MCI unchanged. Alternatively, it can act to manipulate the exchange rate, though it normally does not do so.

The MCI is used as a guide for the Bank's policies. The Bank formulates its expectations on the state of the Canadian economy and those of its major trading partners, decides also on the desired rates of inflation and growth in aggregate demand, and determines the target values of MCI that would achieve these goals. Monetary aggregates, along with other macroeconomic variables, are used as information variables. The Bank does not specify a target path for the MCI, nor set a target path for the exchange rate, nor try to ensure that its actions result in the specific ratio of one-third in interest rate to exchange rate changes. The MCI is used as an operational guide, but the main focus is on its goals defined in terms of aggregate demand and price stability.

The Bank sets the overnight loan rate, with a range of 50 basis points, as its operational target to achieve its desired value of the MCI. It allows the financial institutions and the markets to determine the actual amounts of the monetary aggregates on the basis of the targeted overnight loan rate. Its money market operations are used to hold the overnight rate

24 The actual formula is:

$$\text{MCI} = (\text{CP90} - 7.9) + (100/3)(\ln \text{G10} - \ln 0.8676014)$$

CP90 is the 90-day commercial paper rate and G10 is the Canadian dollar index against G-10 currencies, with its base value for 1981 set at 1.

in the specified range. Movements in the overnight rate in turn induce changes in the other interest rates and the exchange rate.

The Bank of Canada tries to influence the overnight rate through changes in the settlement balances[25] held with it by the direct clearers, mainly the commercial banks, in the Canadian payments system. Positive amounts of these balances do not pay interest, but any negative amounts have to be covered by overdrafts at the bank rate. While such changes in settlement balances can be brought about by open-market operations, the Bank usually relies upon daily transfers of government deposits between it and the direct clearers, making such transfers and the resulting supply of settlement balances its main instrument for changing the monetary base and exercising control over the economy.

The Bank of Canada believes that uncertainty is inimical to the proper functioning of the financial markets and the efficiency of the economy, and that the uncertainty of monetary policy can adversely affect saving and investment in the economy. In an attempt to reduce such uncertainty, changes in the target range for the overnight rate are immediately made known to the public, and the intended course of the monetary policy of the Bank is continuously explained to the public through publications and speeches of the Governor of the Bank and its officials.

11.7 Inflation targeting and the Taylor rule

Currently, the central banks of Britain, Canada, the USA and many other countries are said to follow "inflation targeting," which is the pursuit of an inflation target. However, in reality, they have adopted a Taylor rule (Taylor, 1993; also see Chapters 12 and 13) of the general form:

$$r_t = r_0 + \alpha(y_t - y^{\mathrm{f}}) + \beta(\pi_t - \pi^{\mathrm{T}}) \qquad \alpha, \beta > 0 \tag{1}$$

where r is the *real* interest rate to be set by the central bank, y is real output, y^{f} is full-employment output, π is the actual inflation rate, π^{T} is the inflation target of monetary policy and the subscript t refers to period t. π^{T} is called the *target inflation rate*. Similarly, y^{f} is the *target output level*. $(y_t - y^{\mathrm{f}})$ is (minus of) the output gap. r_0 should be the long-run real interest rate.

The Taylor rule embodies two objectives: stabilizing inflation at its target rate and stabilizing output at its full-employment level. While movements in the inflation rate and output are, on average, positively correlated, they are not necessarily so over short periods. Under this rule, monetary policy raises the interest rate if inflation is above target and output is above its full-employment level. Some proposals for the Taylor rule include other variables, such as the exchange rate.

Note that, in practice, central banks set the nominal interest rate rather than the real one. Consequently, the Taylor rule implies that if $\pi_t - \pi^{\mathrm{T}} > 0$, the nominal rate would rise more than the inflation rate, and if $\pi_t - \pi^{\mathrm{T}} < 0$, the cut in the target real rate would mean that the nominal rate would fall more than the inflation rate. Such a policy is one of "leaning against

25 Settlement balances is the term being currently used to designate the deposits held by designated financial institutions, called *direct clearers* with the Bank of Canada. In the absence of any legally required reserves, these deposits are held voluntarily by the direct clearers and are used to settle their daily imbalances in receipts and payments against each other. This settlement is done through their accounts at the Bank.

the wind." For any given inflation rate, the greater the value of β, the larger will be the change in the real and nominal rates and the stronger the movement to stabilize the inflation rate at its target value.

While monetary policy can use either the interest rate or some monetary aggregate, or some combination of the two, central banks in Britain, Canada and the USA, as well as in many other countries, use the interest rate, rather than a monetary aggregate, as the operating target of monetary policy, though the European Central Bank also monitors movements in monetary aggregates, especially M3, as part of its formulation of monetary policy. The interest rate usually targeted by central banks is the overnight loan rate in the market for reserves, but it could also be the discount/bank rate at which the central bank lends to commercial banks.

A major advantage of specifying the inflation target in the Taylor rule is that it encourages the transparency of monetary policy and anchors inflation expectations, which provides guidance for wage demands, investment plans, etc. Since the 1980s, central banks' achievement of their low inflation targets has increased the credibility of their targets and policies, which has meant that the public's inflationary expectations are now closely determined by the central bank's inflation targets (see Chapter 12 on the credibility of monetary policy).

Empirically, while none of the central banks has announced its form of the Taylor rule or even a commitment to it, this rule has performed quite well in econometric studies for Britain, Canada and the USA, as well as for many other countries. Its use has been credited with a reduction in inflation from high levels during the 1970s and 1980s to the current low levels. It has also been credited with reductions in the volatility of inflation since the 1980s (Clarida *et al*., 1998; Sims, 2000; Moreno and Rey, 2006). Chapters 13 and 15 provide further discussion and references on the Taylor rule.

11.8 Currency boards

Some countries, more so in the past but rarely nowadays, had currency boards instead of central banks.[26] With a currency board, the country maintains a fixed exchange rate against a designated foreign currency, and the monetary base – a liability of the currency board – is backed by its foreign exchange reserves. As these reserves increase – for example, through a balance-of-payments surplus – the currency board increases the monetary base and the money supply in the economy increases. Conversely, as foreign exchange reserves fall, the monetary base and the money supply are decreased. Other than this, the currency board does not have discretion to change the money supply or manage interest rates and therefore cannot pursue domestic monetary policies.

Currency boards were common in the colonies of imperial countries – for example, the UK – during the first half of the twentieth century. They were a means of linking the currency and economies of the colonies to those of the imperial country. Further, if the imperial currency was under the gold standard – that is, with its value fixed in terms of gold – the colonies also indirectly adhered to the gold standard. Such currency boards were usually replaced by central banks on independence. In other cases, countries, though independent, maintained currency boards with a strict adherence to the gold standard, implying a fixed value of the domestic currency in terms of gold.

26 As of 1998, currency boards existed in Hong Kong, Argentina, Estonia and Lithuania and a few other countries.

Note that most of the aspects of this chapter, bearing on the goals, instruments and targets of monetary policy, do not apply to currency boards.

Conclusions

The monetary sector is central to modern economies and its proper functioning is critical to the levels of employment, output and growth. As an illustration of this importance, it is now generally agreed that monetary failure caused, or was a major main contributor to, the Great Depression of the 1930s (Friedman and Schwartz, 1963).

The central bank is the custodian of the health, efficiency and performance of the financial sector. The monetary policy pursued by the central bank is therefore fundamental to the performance of the economy. This chapter has shown that the goals and the favored instruments of central banks have varied over time and differ among countries. Currently, the dominant belief of most central banks is that they can best promote output, employment and growth through fairly stable prices. In particular, the belief is that continual or discretionary attempts to increase aggregate demand through monetary policy do not yield higher output or reduce business fluctuations over time. Further, monetary policy acts on the economy with a sufficient lag so that, with the future course of the economy being difficult to predict accurately, the proper formulation of monetary policy is an art and, as such, potentially susceptible to error.

In the 1990s the dominant goal, pursued in the European Union, Canada, New Zealand, the UK, the USA, and many other countries, was that of price stability, encompassed in the notion of inflation targeting, which is the maintenance of a low rate of inflation (about 1 to 3 percent). Given the long-run neutrality of money, this goal is sometimes conveyed as an ultimate goal and at other times as an intermediate goal towards improving the growth rates of output and employment in the economy. For Canada, the European Union, New Zealand and the UK the target for the inflation rate is explicitly announced, whereas the USA does not announce – and therefore, does not give an explicit pre-commitment to – a pre-specified inflation target. However, empirical studies have shown that the Taylor rule (see Chapters 13 and 15) performs quite well for all these countries, so that the monetary policies of the central banks were not aimed exclusively at an inflation target but rather at the deviations of output and inflation from their desired long-run levels.

The use of interest rates as operational targets is to be distinguished from a policy of fixing them administratively for long periods and at relatively low levels, with the ostensible purpose of promoting growth. This was done in many LDCs during the 1970s and 1980s, though there is currently a trend away from this practice. The interest rate is the opportunity cost of loans and should, for the proper allocation of funds in the economy, be determined by competitive forces in the open markets. Setting them or setting artificially low ceilings on them by administrative action introduces inefficiencies into the financial structure of the economy, inimical to the optimal generation of savings, as well as to their allocation by the financial institutions optimally to investment among the sectors of the economy.[27] LDCs are especially prone to setting artificially low interest rates on loans in the organized

27 The continual revision of such ceilings in the light of changes in the market interest rates, as is done in some countries, does moderate this argument. But since they are invariably delays and rigidities in changing administered interest rates, such rates almost always introduce some inefficiency in the financial sectors of the economy.

financial sector. This could also happen in developed economies in which the interest rate is used as an operating target.[28]

Summary of critical conclusions

- Historically, most central banks have had the mandate to pursue a number of macroeconomic goals, including price stability, low unemployment and high growth. Achievement of multiple goals is only possible if the economy does allow such a possibility and the policy maker has enough policy tools.
- Since the early 1990s, many economists have recommended – and many central banks have followed – inflation targeting. However, the Taylor rule, embodying a trade-off between deviations of inflation and output from their desired long-run levels, better describes the current pursuit of monetary policies.
- While the interest rate was historically the operating target of monetary policy, a diversion to monetary targeting occurred during the late 1970s under the impact of St Louis monetarism. This experiment was not considered to be a success in most countries.
- Most Western countries have reduced percentage reserve requirements on commercial banks to levels that are close to zero, which eliminates changes in these requirements as a tool of monetary policy.
- Since the early 1990s, the most common tools of monetary policy in developed economies have been changes in interest rates, supported by changes in the money supply induced by open-market operations and borrowing from the central bank.

Review and discussion questions

1. Historically, what goals were mandated for central banks? Why have the goals pursued in recent years been narrowed to "inflation targeting"?
2. Can central banks pursue and achieve multiple goals or must they be confined solely to fighting inflation? What goals are embedded in the Taylor rule? Discuss.
3. What is the lender-of-last-resort function of the central bank in modern economies? What is its justification? Should commercial bank borrowing from the central bank be a privilege or a right? Discuss.
4. Should the United States and Canada follow the example of Britain in giving the government the power to set the ultimate goal or goals of monetary policy, while leaving the implementation of these to the central bank? Or should Britain follow the example of the United States and Canada on this issue? Discuss.
5. What are the tools available to the central bank for controlling the money supply? Discuss how manipulation of each of these tools will change the money supply and how reliable each tool is likely to be.
6. Why has the use of changes in reserve requirements as a tool of monetary policy been largely abandoned in Western economies? What were the reasons for the virtual elimination of reserve requirements? Is there a case for their revival and usage as a tool of monetary policy in the context of the country you live in? In LDCs?

28 Many economists contend that the Fed maintained interest rates too low during 2002–2007, thereby contributing to high aggregate demand – as well as to the housing price bubble during these years and its collapse in 2007.

7. How can central bank discounting cause procyclical movements in the money supply? How can the central bank eliminate such a movement? Discuss.
8. The pursuit of selective fiscal policies in the form of tax exemptions and subsidies is common in almost all countries, whereas the usage of monetary policy on a selective basis is rare in the financially developed economies. Why? Should the use of selective monetary policies be abandoned for the LDCs also?
9. Suppose that for a given economy the preconditions for the effective pursuit of open market conditions are not met. What monetary policy tools are likely to be the most effective ones for such an economy? Relate your recommended tools to the preconditions that are not met.
10. What does financial intermediation mean? What are the different financial intermediaries in your country. Which of their liabilities would you classify as components of the monetary aggregates? Discuss.
11. Suppose that instead of imposing reserve requirements on demand deposits in the commercial banks, the central bank does the following: require automobile owners to hold $500 of non-interest-bearing deposits with the central bank.
    i. How would the determination of the price level differ between the two arrangements?
    ii. How would the real consequences of the two arrangements differ?
    iii. Are there any special characteristics of demand deposits or any other reasons that make one of these arrangements preferable to the other?
12. "Due to fairly radical changes in the structure of the financial sector in recent decades, the importance of demand deposits has declined considerably and the role of commercial banks has changed dramatically. The main stream of monetary theory and practice continues to be directed mainly at demand deposits and commercial banks as purveyors of checking accounts. But checking deposits are simply one kind of liquid asset and banks are simply one kind of intermediary. Therefore, to single out checking deposits and commercial banks for special analytical treatment is mistaken. Effective control of the economy by the central banks requires control over all types of liquid assets and over the liabilities of all financial intermediaries." Discuss.
13. "The standard practice of governments whereby they define the monetary unit is unnecessary and undesirable. The private sector should be encouraged to choose its own standards in a free competitive market" (Friedrich Hayek). Discuss.

References

Clarida, R., Galí, J. and Gertler, M. "Monetary policy rules in practise: some international evidence." *European Economic Review*, 1998, 42, pp. 1033–67.

Friedman, M. "Nobel prize lecture: inflation and unemployment." *Journal of Political Economy*, 85, 1977, pp. 451–73.

Friedman, M. and Schwartz, A.J. *A Monetary History of the United States*. Princeton, NJ: Princeton University Press, 1963.

Goodhart, C.A.E. "Central banking." *The New Palgrave Dictionary of Economics: Money*. London: Macmillan, 1989.

Goodhart, C.A.E. *The Central Bank and the Financial System*. Cambridge, MA: MIT Press, 1995.

Moreno, A. and Rey, L. "Inflation targeting in Western Europe." *Topics in Macroeconomics*. Berkeley Electronic Press, 6, 2006, Article 6.

Selgin, G. *The Theory of Free Banking: Money Supply under Competitive Note Issue*. Totowa, NJ: Rowman and Littlefield, and the Cato Institute, 1988.

Sims, C.A. "A review of monetary policy rules." *Journal of Economic Literature*, 39, 2000, pp. 562–6.

Taylor, J.B. "Discretion versus policy rules in practise." *Carnegie–Rochester Conference Series on Public Policy*, 39, 1993, pp. 195–214.

12 The central bank

Independence, time consistency and credibility

This chapter focuses on the analytical treatment of three major issues: independence of the central bank, time consistency of policies and the credibility of central bank objectives and policies. Assuming a potential for tradeoffs among goals, this chapter examines the determination of the choices made and the potential for conflicts among the monetary and fiscal authorities. This discussion leads to the examination of the independence of central banks from governments and legislatures.

This chapter also shows the superiority of intertemporal optimization policies to myopic ones, which can have an inflationary bias. Intertemporal optimization over time provides two types of policy approaches. One of these is the time-consistent one in which the policy path for the current and future periods is derived only once and followed for all future periods. The second approach allows reoptimization every period with an unchanging objective function.

Finally, the chapter investigates the superiority of prior commitment to future objectives and policies, and credibility of the central bank.

Key concepts introduced in this chapter

- Preferences over goals
- Economy's constraints on the tradeoff among goals
- Conflicts among policy makers
- Central bank independence
- Myopic optimization
- Time consistency of policies
- Reoptimization with an unchanging objective function
- Credibility of policy and commitment

This chapter focuses on several issues important to the formulation of monetary policy by the central bank. Among these are the choice among goals and the possibility of conflicts between the monetary and fiscal authorities in the attainment of their desired targets and goals. In cases of such conflicts, the ability of the central bank to pursue its own choices becomes important and is discussed under the heading of the independence of the central bank.

The other two major issues addressed in this chapter are those of the time consistency and credibility of policies. The time consistency of monetary policies deals with the question of whether the central bank should determine its policies for the future periods within its horizon and stick to them, or should retain discretion to reformulate its policies

as time passes. Discretionary policies can be arbitrary ones, be derived from one-period ("myopic") optimization or be based on continual intertemporal reoptimization as each period passes. Related to the issue of time consistency of policy is the important one of maintaining the credibility of the central bank among the public and the consequences of a failure to maintain credibility.

Section 12.1 presents the optimizing framework for making choices among goals and targets. This analysis follows the utility maximization approach, subject to the constraints on targets set by the economy. Section 12.2 considers the case of two policy makers, the central bank in charge of monetary policy and the government in charge of fiscal policy, and shows that there is considerable scope for conflict in the pursuit of these policies. Section 12.3 addresses the important issue of the independence of the central bank from the government. Section 12.4 presents the analyses of policies based on myopic optimization, intertemporal time consistent optimization and intertemporal reoptimization. Finally, Section 12.5 examines the credibility of central bank policies and their effectiveness, as well as the impact on credibility of gradualist versus cold-turkey attempts to lower the rate of inflation.

12.1 Choosing among multiple goals

As discussed in the last chapter, economic theory and central bank beliefs prior to the 1980s had indicated that several goals could be addressed through monetary policy. For the analysis of such a possibility, this section assumes that the central bank has a multiplicity of goals. Focusing only on the primary goals, the instruments available for achieving the multiple goals are severely limited in number and scope so that not all the goals can be attained through the use of monetary policy. Therefore, the central bank has to make a choice among its desired goals or combinations of them.

Assume that the central bank's preferences over the goal variables are consistent and transitive, so that there exists an ordinal utility function over the goal variables. For diagrammatic analysis, the indifference curves between any given pair of these variables can be derived from this utility function.

Choosing between inflation and unemployment

The goal variables of many central banks include the rate of inflation and the unemployment rate, which can be a goal variable in its own right or a proxy for the output gap. Assume that the central bank's preferences over these variables can be encompassed in an objective/utility function of the form:

$$U = U(\pi, u) \tag{1}$$

where π is the rate of inflation, u is the rate of unemployment, and $U_\pi, U_u < 0$. Hence, the indifference curves in the (π, u) space are negatively sloped. Further, it is reasonable to assume that the undesirability – that is, disutility – of each variable keeps on increasing, *ceteris paribus*, with higher levels of it, so that $U_{\pi\pi}, U_{uu} < 0$. Hence, as the rate of inflation rises, the central bank is willing to accept a higher marginal increase in the unemployment rate in order to prevent a further rise in the rate of inflation, so that the indifference or tradeoff curves between the rates of inflation have the usual convex shape, as shown in Figure 12.1 by the curves I^{CB} and I'^{CB}. A host of such curves exist, with a curve passing through every

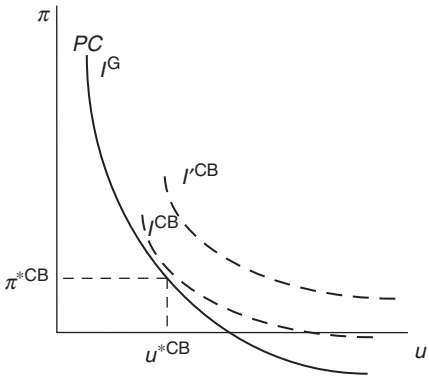


Figure 12.1

point in the quadrant. Note that since being on a lower curve is preferred to being on a higher one, the central bank will want to be on the lowest attainable indifference curve.

A common variant of the above objective function is:

$$U = U(\pi - \pi^*, y - y^{\mathrm{f}}) \qquad U_{\pi\#}, U_{y\#} < 0 \tag{2}$$

where $\pi^{\#} = \pi - \pi^*$, $y^{\#} = y - y^{\mathrm{f}}$, so that $\pi^{\#}$ and $y^{\#}$ are the respective gaps between the actual and the desired values of π and y. Since the central bank's choices are limited by constraints imposed by the economy on the values of π and u that can be attained, the central bank's decision problem is optimization of utility subject to these constraints.

The preceding discussion has been in terms of a general utility function. Current monetary and macroeconomic theory prefers to use an intertemporal, as against a myopic (one period) utility function. The choice between intertemporal and myopic optimization is discussed later in this chapter in the context of the time consistency debate. Chapter 15 re-visits these issues in the discussion on the Phillips and new Keynesian curves.

General analysis of the choice among goals when the economy allows a tradeoff

The preceding utility function can be generalized to the case of n variables. The formal analysis of this general case assumes that the central bank has a utility function:

$$U = U(x_1, \ldots, x_n) \tag{3}$$

At this stage, we assume that the central bank's choice is subject to only one constraint, which has the form:

$$f(x_1, \ldots, x_n; \boldsymbol{z}, \boldsymbol{\Psi}) = 0 \tag{4}$$

where:

x_i = ith goal variable
$\boldsymbol{z}$ = vector of instruments available to the central bank
$\boldsymbol{\Psi}$ = vector of exogenous variables.

The goal variables can be levels of variables, their growth rates or even such variables as the output gap and the "inflation gap," where the gaps are deviations from their desired levels. $U(.)$ represents the central bank's preferences over its goals. These depend upon the organizational structure of the central bank, the interactions between the policy makers, their perceptions of society's goals, the structure of the economy and of what is achievable, political pressures, and so on. Equation (4) should properly specify the actual form of the constraint. However, this form is usually not known, so that the form of the constraint used by the central bank is that perceived by it and is based on its knowledge of the structure of the economy and the political and social environment. However, under imperfect information on the economy, the relevant constraint perceived by the central bank may not necessarily or even usually be the actual one imposed by the economy.[1] The central bank is taken to maximize (3) subject to (4) in order to determine its optimal choices among the goals.

The basic objections to the use of a preference function are to its requirements of consistency and transitivity in making choices. The central bank's decisions are made by a host of individuals and its major choices, if consciously made, are by a group. Such group decisions in a democratic framework need not necessarily be consistent and transitive, even at a point in time, let alone over time. Further, the policy makers in the central bank change over time, so that its preferences are likely to shift over time. Hence, one must be cautious in explaining the choices among goals made by the central bank within a static utility function and its implied set of indifference curves, especially when there has been a change in its management.

In spite of these objections, the preceding analysis furnishes considerable insights into the problem of choice among alternative goals. Empirical and descriptive studies using data up to the 1980s indicate considerable validity for this analysis and show that the central banks often manipulated their policy instruments, such as the monetary base and interest rates, in a systematic fashion to address their chosen goal levels.

Choices under the economy's supply constraint

There are several forms of the economy's supply constraint relating u and π. Of these, the Phillips curve (see Chapter 15) was proposed in 1958 by A.W. Phillips and soon began to be treated as the economy's constraint between inflation and unemployment. Its general form was:

$$u = f(\pi) \qquad f' < 0 \tag{5}$$

This constraint allows the central bank to trade between higher inflation and lower unemployment. Optimization of the utility function (1) subject to (5) yields the optimal values of π and u, with higher inflation rates yielding lower unemployment.

While most Keynesians of the 1960s and 1970s accepted some form of (5) as the economy's constraint and explained central banks' monetary policy under it, many economists, especially neoclassical ones, believed that the economy had a vertical Phillips curve for the long run. Their arguments were subsequently refined by Friedman and Lucas (see Chapters 8 and 14)

1 Therefore, there is ample scope for wide differences among political parties, economists, political scientists and the public, etc., on what they believe can be accomplished by any given set of policies, which may differ from what in fact can be accomplished.

who asserted that the proper form of the economy's constraint was the Friedman–Lucas aggregate supply curve, also known as the expectations-augmented Phillips curve. This constraint, under the assumption of rational expectations, is of the form:

$$(u - u_n) = f(\pi - E\pi) \qquad f' < 0 \tag{6}$$

where $du/dE\pi = 0$ and the use of a systematic monetary policy by the central bank changes both π and $E\pi$ by the same extent, so that $(\pi - E\pi)$ would not change. This constraint belongs to the modern classical approach (see Chapters 1 and 14). According to it, unemployment can only be made to deviate from its natural rate through unanticipated inflation, which, given rational expectations, requires a random monetary policy. Therefore, there is no tradeoff between these variables which systematic monetary policy by the central bank can exploit, so that the recommendation is that the central bank should adopt the target of price stability. Many central banks adopted this economic framework in the 1990s and some economists advocated that price stability or a low rate of inflation should be the central bank's *only* goal variable. Further, in the 1990s, many central banks came to believe that higher rates of inflation have little to contribute in terms of higher output, while they could lead to escalating inflation and lower output. At the same time, negative rates of inflation are considered inimical to full employment because of their potential for causing a recession. Given these beliefs, the utility-maximizing goal rate of inflation would be zero, or a low positive inflation rate.

As pointed out earlier, some economists choose to work with the objective function:

$$U = U(\pi - \pi^*, y - y^{\mathrm{f}}) \qquad U_{\pi\#}, U_{y\#} < 0 \tag{7}$$

where $\pi^{\#} = \pi - \pi^*, y^{\#} = y - y^{\mathrm{f}}$, so that $\pi^{\#}$ and $y^{\#}$ are the respective gaps between the actual and the desired values of π and y. Given this objective function, maintaining the expectations-augmented Phillips curve as the constraint, as well as assuming that the central bank does not want to or cannot fool the public by causing unanticipated inflation, implies that the optimal values of inflation and output are π^* and y^{f}, of which the former can be achieved through systematic monetary policy. The latter is not affected by monetary policy but is determined by the long-run performance of the economy.

However, for imperfect competition and price rigidities, the new Keynesians propose a different form of the supply constraint (see Chapter 15) to (6), so that their policy recommendations would differ from the above. This form is known as the new Keynesian Phillips curve.

The preceding discussion of the Phillips curve, its expectations-augmented version and the new Keynesian Phillips curve illustrates that the accurate form of the economy's constraint is usually not known. There are continual disputes among economists even about its general form, less alone the specific one with numerically specified values of the parameters.

12.2 Conflicts among policy makers: theoretical analysis

Another application of the utility approach is to the choices exercised by several (at least two) policy makers over the same set of goal variables. Different policy-making bodies in the economy are likely to have different preference functions and hence different indifference

curves between any given pair of variables. Therefore, the formal optimization analysis for two policy makers A and B would be:

(1) For policy maker A:

$$\text{Maximize} \qquad U^{\mathrm{A}} = U^{\mathrm{A}}(x_1, \ldots, x_n) \tag{8}$$

subject to A's perceived constraint:

$$f^{\mathrm{A}}(x_1, \ldots, x_n; \boldsymbol{z}, \boldsymbol{\Psi}) = 0 \tag{9}$$

(2) For policy maker B:

$$\text{Maximize} \qquad U^{\mathrm{B}} = U^{\mathrm{B}}(x_1, \ldots, x_n) \tag{10}$$

subject to B's perceived constraint:

$$f^{\mathrm{B}}(x_1, \ldots, x_n; \boldsymbol{z}, \boldsymbol{\Psi}) = 0 \tag{11}$$

The superscripts A and B refer to the policy maker. Since both the utility functions and the perceived constraints can differ, the optimal values of the goals for $x_1^{\mathrm{A}}, \ldots, x_n^{\mathrm{A}}$ will differ from $x_1^{\mathrm{B}}, \ldots, x_n^{\mathrm{B}}$, so that working at cross purposes can be a common phenomenon, rather than a rare occurrence, among policy makers in the economy. This possibility depends upon the differences in the utility functions, becoming reinforced by any differences in the policy makers' perceptions of the actual present and expected course of the economy. In most cases, such conflicts in the understanding of the economy and desirable tradeoffs among objectives by the fiscal and monetary authorities of a given country tend to be mild. However, they can erupt into open and sometimes acrimonious public debate in times of radical economic and political change and of differences in ideology.

The two principal tools for the control of aggregate demand are monetary and fiscal policies. In a country with an independent central bank, monetary policy is in the control of the central bank whereas fiscal policy is in the hands of the legislature and the government. The latter depends on the public for electoral support and generally tends to attach greater undesirability to increases in unemployment relative to increases in inflation than does the central bank which is usually more vitally concerned with inflation. Formally, in terms of the marginal rates of substitution of the two policy makers, $\partial\pi/\partial u^{\mathrm{CB}} < (\partial\pi/\partial u)^{\mathrm{G}}$, where CB stands for the central bank and G for the government, implying (Figure 12.2) that the indifference curves of the central bank are steeper than those of the government. This implies that for a given constraint $f(\pi, u) = 0$, the central bank would adopt a monetary policy aimed at achieving a lower rate of inflation than the government. This is illustrated in Figure 12.2 in which the central bank's indifference curves are shown by I^{CB}, the government's by I^{G}, and the economy's (common) constraint is shown by *PC*. The central bank's optimal choice is for $(\pi^{*\mathrm{CB}}, u^{*\mathrm{CB}})$ and the government's is for $(\pi^{*\mathrm{G}}, u^{*\mathrm{G}})$, implying a more expansive stance by the government for the economy relative to that by the central bank. There therefore exists in this case a conflict between the central bank and the government on the desired rates of inflation and unemployment for the economy. If each tries to achieve its goals through the policy at its command, neither will achieve their own goals.

Over time, the political process may bring about a narrow "consensus range" within which the differences between the central bank and the government over the desired goals are

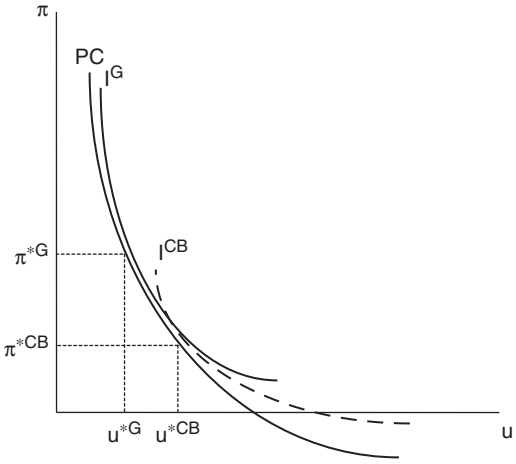


Figure 12.2

mild and accommodation is made easily. But a sharp change in the course of the economy outside such a range, or a sharp change in the objective functions of one of the parties to the process, as after an election that brings a new political party with a different ideology to power, may provoke an open conflict between the policy makers which takes time to resolve.

The potential for conflicts between two independent policy makers leads to strategic considerations where each "player" tries to outsmart the other. The theoretical analysis appropriate to such interactions belongs to game theory. Such analysis is outside the scope of this book. A review of it is provided in Blackburn and Christensen (1989).

12.3 Independence of the central bank

As shown by the preceding analysis and examples, potential conflicts are inherent in a situation where the central bank is free to formulate monetary policy independently of the government, which is in charge of the fiscal policies and the management of the public debt. This conflict can be about the ultimate goals of full employment and price stability. However, it is more often about intermediate targets, such as the desirable levels of interest rates or exchange rates, or because of the introduction of other ancillary objectives such as the costs of servicing the public debt or financing fiscal deficits.

While the potential for conflict can be avoided by the subordination of the central bank to the government, its independence usually ensures lower inflation rates. As discussed in Chapter 11, the USA Fed is now one of the most independent central banks in the world. In practice, the Bank of Canada has retained its independence, though there is close consultation between the Governor and the Minister of Finance on inflation targets and policy changes. The British experiments with its central bank's independence have varied over time.[2] While the

2 Goodhart (1994) provides a compact statement of the issues related to central bank independence, time consistency and credibility, especially from the perspective of the British experience.

Bank of England was historically a quasi-private bank, independent of the government, the dominance of the government over the Bank of England was legislated by the Bank Act of 1946, which nationalized its ownership. It also allocated the choice over the goals and targets of monetary policy to the Chancellor of the Exchequer, representing the government, leaving the Bank of England with a consultative and implemental role. In 1997, the Bank of England and the Monetary Policy Committee were given operational independence toward the achievement of the inflation target set by the Chancellor. These arrangements were enacted into law by the Bank of England Act of 1998. Under it, the Chancellor sets the goals of monetary policy – currently a target value of the inflation rate – and the Bank and the Monetary Policy Committee have the responsibility of formulating and implementing policies for their achievement. Therefore, as of 1999, Britain allocates only operational but not goal independence to its central bank. At a formal level, this degree of independence differs from that of the central banks of the Canada and the USA, which possess both goal and operational independence. The Bank of England is now one of the national banks within the federal structure of the European System of Central Banks. The European Central Bank (ECB) is, by legislation, independent of the government of the European Union.

In many other countries, including many LDCs, the subordination of the central bank to the fiscal authorities is fairly common in practice, even if the legislation formally makes the central bank independent of the government.

Such lack of independence of the central bank in the formulation and pursuit of monetary policy represents a threat to the pursuit of appropriate monetary policies for price stability. In fact, many empirical studies for the 1970s and 1980s showed that countries that retained actual independence of the central bank from the government tended to have lower rates of inflation; conversely, the lack of independence resulted in higher inflation rates (Alesina and Summers, 1993).

The trend since the 1980s has been strongly towards granting greater or complete independence. By now, almost all developed economies have independent central banks.

Development strategies in LDCs, financing fiscal deficits and central bank independence

The issue of central bank independence takes on another dimension in countries that incur large and persistent deficits but do not possess adequate capital markets to finance them through new issues of public debt, and need the central bank to do so through an expansion of the monetary base. This often happens during wars, even in the developed economies, but has occurred most noticeably in recent decades in the LDCs.

LDCs tend to have low output per capita and are not able to raise adequate tax revenue for their desired levels of public expenditures. The latter are in many countries swollen by their plans for public development projects or the deficits of their public sector undertakings. Further, their domestic financial markets are under-developed and cannot support much, if any, government borrowing, and their ability to borrow abroad is also severely limited. As a result, many LDCs resort to increases in the monetary base, either directly or indirectly through the compulsory sale of government bonds to the central bank. This process requires the subservience of the central bank and its policies to the fiscal needs of the government, and destroys the central bank's independent control over monetary policy.

Whether such an arrangement is advantageous to the economy is in considerable doubt. On the positive side is the financing of public projects that would otherwise not have been

financed, or provision for social objectives such as health, education and alleviation of poverty. On the negative side is the subordination of increases in the monetary base to budget deficits, with the consequent loss in control by the central bank over monetary policy and over aggregate demand and inflation in the economy. From the perspective of price stability, this loss of independent control of the monetary base by the central bank severely limits its capacity to control inflation. Many empirical studies have documented that countries with independent central banks, which are not necessarily obliged to finance the budget deficits, tend to have lower rates of inflation. Another negative aspect of this arrangement is that the borrowing for the public projects thus financed is not done in competitive markets at market-determined rates. Hence, allocative efficiency suffers and private projects that could be more efficient and could have been undertaken are crowded out. These efficiency losses could be considerable and, in the opinion of many economists, have contributed to the low growth of those LDCs that resorted heavily in the past to the financing of governmental deficits by increases in the monetary base.

While few central banks in developing countries were effectively independent of the government prior to the 1990s, the economic arguments and evidence supporting independence have since then led many developing countries to grant much greater independence to their central banks.

Central bank independence in practice

A country can have laws formally legislating the independence of its central bank from the government. However, its independence in practice will usually also depend on additional factors, such as the mode of appointment of the bank's directors, the duration and terms of their appointment, their relationship with those in the legislature and the government, power politics, and how the spirit of the law is respected in the country in question. On this point, Cukierman *et al*. (1993) reported that while legal independence was an important determinant of inflation in the industrialized economies, with a negative coefficient, the rate of inflation was strongly and positively related to the central bank governor's rate of turnover, rather than to legal independence, in the developing economies.

Hence, legal independence does not always ensure factual independence. In general, the factual independence of the central bank from the government normally depends on its acceptance by the main political parties of the country and substantial support for it among the public. This acceptance, in turn, depends on the rule of law in the country, the belief in the integrity and commitment of the central bank directors and governors and the past record of the central bank, as well as the transparency of its policies and its accountability.[3]

Goodhart (1994) finds that central bank independence, if effective, is a way of making its commitment to price stability more credible and makes the achievement of low inflation more likely. But he finds no evidence that it lowers unemployment or the "loss ratio," defined as "the number of extra man-years of unemployment necessary to lower inflation by one per cent" (p. 68). Eijffinger and de Hahn (1996) concluded from their review of empirical studies that the weight of empirical evidence does support the claim that independence of the central bank from the government does reduce inflation. However, such independence does not

3 Eijffinger and de Hahn (1996) provide a thorough review of the issues and empirical studies on central bank independence.

seem to have increased employment and growth. Further, there were higher rather than lower disinflation costs associated with independent central banks.

12.4 Time consistency of policies

The proper design of monetary policies[4] over time requires that the central bank have an intertemporal objective function to rank alternative policies and know the economy's constraints, as well as possess knowledge of the current and future responses of the economy to its policies and how it intends to set its policies in the future.

A *time-consistent policy path* is one that is derived from optimizing an intertemporal objective function subject to the appropriate constraints describing the behavior of the economy, with the optimal policy path over time derived *once-for-all* and followed as time passes. The latter requires a commitment to pursue the derived set of policies, both in the current and future periods, so that the central bank would have to resist the temptation and the political pressures to deviate from this path. Hence, time consistency of policies is related to the issues of the independence of the central bank and of a commitment regime under which the central bank will maintain a pre-set future policy path.

To illustrate, suppose that the central bank does want to pursue time-consistent policies with the long-term goal of price stability, while the government has the objective of getting re-elected which, given a short-run Phillips curve tradeoff between inflation and unemployment, is enhanced by short-term inflationary monetary policies. The central bank is more likely to resist pressure from the government if it is independent of the political process. Therefore, the independence of the central bank from the government and the legislature improves its ability to pursue time-consistent policies. This reasoning is now widely accepted, so that, as pointed out earlier, the central banks of most developed countries and of many emerging ones now possess a high degree of independence of the government.

Time-consistent policies are generally compared with discretionary policies. *Discretionary policies* allow the central bank discretion to deviate or not from the pre-set policy path. Under them, the central bank retains the right to pursue policies as it thinks fit at the time the policies are pursued. The set of policy types is:

1 *Purely arbitrary policies*.
2 *Myopic (one period) policies*. These are derived from myopic (short-term, usually one period) optimization with short-term goals subject to the constraints for the current period only and with expectations taken as exogenously given. Further, under this procedure, the policy maker does not give any advance commitment, even on maintaining the objective function, on its future policies.
3 *(Intertemporal) reoptimization policies*. These policies are derived from intertemporal reoptimization *each period* for that period and future ones, with an unchanged intertemporal objective function, which is maximized subject to the long-run or multi-period constraints specified by the structure of the economy. Under this procedure, the policy maker gives a commitment to maintain the same intertemporal objective function over time, but the relevant constraints can shift with the passage of time. The optimal

4 Fischer (1990) provides an excellent review of the literature on different types of policies, their evolution and foundation, as well as a comparison of time-consistent versus discretionary policies.

policy is followed only for the optimizing period, since the following period's policy will be the outcome of that period's optimization process. Policies of this type are labeled "reoptimization policies" in this chapter and are the outcome of a dynamic reoptimization process performed each period, but the optimal policy is followed for the optimizing period only.

4 *(Intertemporal) time-consistent policies*. These policies are derived from *once-for-all* intertemporal optimization, with an intertemporal objective function over goals, maximized subject to the long-run or multi-period constraints specified by the structure of the economy. The policy path, once derived, is followed in the initial (i.e. optimization period) and all future periods, so that the optimization exercise is done only in the initial period. If the initial period has already passed, optimization is not undertaken in the current period. There is clearly an implicit or explicit commitment to stick to the policy path derived in the initial period.

Compared with the reoptimization procedure, which requires reoptimization each period, time-consistent policies are derived from just one optimization. Note also that the time-consistent optimal policy path (i.e. under once-for-all optimization) does not imply unchanged or identical policies for each period since the period constraints can differ, as, for instance, because of foreseen business cycle fluctuations.

Policies of types 1 to 3 are usually classified as discretionary since the policy maker does not make a prior commitment to following pre-announced policies for future periods. However, note that "reoptimization policies" are not arbitrary or myopic, and are discretionary only in the very limited sense that the policy maker changes the policy pursued from one period to the next only if the intertemporal reoptimization, with an unchanged objective function, implies such a change.

Contributions in the 1970s and 1980s on the proper design of policies showed that policies conducted on an arbitrary or myopic basis generally result in poor long-term outcomes. In particular, relying on a one-period Phillips curve tradeoff between output or unemployment and inflation, expansionary policies to boost output above its sustainable level would not keep output on average above its long-run level but would generate inflation, possibly accelerating inflation, on a continuing basis. This result is known as the inflationary bias of myopic, discretionary policies. Over time, this realization would sooner or later lead to the reversal of such policies, so that they would be "time inconsistent."

Time-consistent and reoptimization policies are clearly preferable to arbitrary policies. From the perspective of sustainable long-run goals, they are also preferable to myopic policies. However, it is not clear whether time-consistent policies are also superior to reoptimization policies. Offhand, the intuitive presumption is that the reoptimization policy procedure is preferable since it maintains continuous policy flexibility and since, with reoptimization at the beginning of each period, it eliminates from decision-making what is gone and past – a procedure common in economics. However, this intuitive presumption was called into question by Kydland and Prescott (1977). The discussion below is based on their analysis on the optimality of time-consistent policies relative to reoptimization policies.

12.4.1 Time-consistent policy path

To reiterate, a time-consistent policy path is one that maintains the a priori determined time pattern of policies. This pre-determined policy pattern involves the derivation of a policy set

at the opening of the initial period under intertemporal optimization over *all* periods and its pursuit as time passes, so that it represents a fixed policy path over time. The predetermined optimal policies may, and often are likely to, differ for different periods since each period has its distinctive elements.

The following analysis assumes certainty, so that the extension of its implications to the real-world uncertainty scenario will have to be examined later. For this analysis, assume a *fixed two-period horizon* and let the objective function of the policy maker over the two periods, $t = 1, 2$, be:

$$U = U(x_1, x_2, \pi_1, \pi_2) \tag{12}$$

where:

U = policy maker's preference function over policies
x_t = economic agents' and the economy's response in period t
π_t = policy variable for period t.

Economic agents are assumed to take account of the policies pursued in making their decisions, so that the response of economic agents and the economy depends on the policies pursued. The response functions are:

$$x_1 = x_1(\pi_1, \pi_2) \tag{13}$$

$$x_2 = x_2(x_1, \pi_1, \pi_2) \tag{14}$$

Note that since this model does not allow uncertainty, the future values of the variables are known at the beginning of period 1. We assume that there are no other constraints to be taken into consideration.

For the intertemporal optimization of the consistent policy path, both policy decisions π_1, π_2 are made in period 1. To derive these decisions, substitute (13) and (14) in (12). This gives:

$$U(.) = U(x_1(\pi_1, \pi_2), x_2(x_1(\pi_1, \pi_2), \pi_1, \pi_2), \pi_1, \pi_2) \tag{15}$$

For the optimal values of π_1 and π_2, the policy maker maximizes $U(.)$ in (15) with respect to π_1 and π_2, so that the two Euler conditions can be solved for the optimal values $\pi_1{}^*$ and $\pi_2{}^*$. These policies are derived from a priori optimization, so that they specify the time-consistent policy path, as defined above. In general, $\pi_2{}^*$ can differ from π_1^*, so that the time-consistent policy path need not have time-invariant (i.e. identical) policies.

However, for the comparison below with reoptimization policies, we need only to consider the first-order Euler condition with respect to π_2, which is:

$$\left[\frac{\partial U}{\partial x_1} \cdot \frac{\partial x_1}{\partial \pi_2} + \frac{\partial U}{\partial x_2} \cdot \frac{\partial x_2}{\partial x_1} \cdot \frac{\partial x_1}{\partial \pi_2}\right] + \left[\frac{\partial U}{\partial x_2} \cdot \frac{\partial x_2}{\partial \pi_2} + \frac{\partial U}{\partial \pi_2}\right] = 0 \tag{16}$$

This simplifies to:

$$\frac{\partial U}{\partial x_2} \cdot \frac{\partial x_2}{\partial \pi_2} + \frac{\partial U}{\partial \pi_2} + \frac{\partial x_1}{\partial \pi_2}\left[\frac{\partial U}{\partial x_1} + \frac{\partial U}{\partial x_2} \cdot \frac{\partial x_2}{\partial x_1}\right] = 0 \tag{17}$$

12.4.2 Reoptimization policy path

Under the reoptimization procedure, the change in the information set from one period to the next one can be of two types:

- Change in information resulting from the mere passage of time, since with the passage of a period the values of the variables in that period have pre-determined values in future optimization.
- Change in information due to shifts in the perceived probabilities and outcomes, either resulting from changes in knowledge or from shifts in the nature of the economy.

While the latter types of shift do routinely occur, we will for the time being bar them from consideration in the current comparison of time-consistent versus reoptimization policy paths. Therefore, in the current comparison, the definition of a reoptimization policy path is limited to the consideration of changes in policies resulting from an information shift (not from a change in knowledge or an unanticipated shock to the structure of the economy) that occurs naturally due to the mere passage of time.

For the reoptimization policy path, the policy maker pursues $\pi^*{}_1$ in period 1, as derived above. However, in period 2, the passage of time has relegated π_1 and x_1 to the past, so that the constraints change. Under the discretionary (now myopic one-period) policy, the decision made in period 2 is based on the optimization with respect to π_2 of the same utility function as in period 1 but now subject to the constraints applicable at the beginning of period 2. That is, optimization in period 2 requires:

$$\max U = U(x_1, x_2, \pi_1, \pi_2) \tag{18}$$

subject to:

$$x_2 = x_2(x_1, \pi_1, \pi_2) \tag{19}$$

$$x_1 = \underline{x}_1 \tag{20}$$

and

$$\pi_1 = \underline{\pi}_1 \tag{21}$$

where the underlining indicates given values. Note that the original first-period constraint (13) is no longer relevant. Substituting (19) to (21) in (18) gives the reoptimization problem as the maximization of:

$$U(.) = U(\underline{x}_1, x_2(\underline{x}_1, \underline{\pi}_1, \pi_2), \underline{\pi}_1, \pi_2) \tag{22}$$

The first-order Euler condition for (22), with utility maximization subject to the only remaining policy variable π_2, is:

$$\frac{\partial U}{\partial x_2} \cdot \frac{\partial x_2}{\partial \pi_2} + \frac{\partial U}{\partial \pi_2} = 0 \tag{23}$$

Let the solution to (23) be π_2^{**}. The discretionary myopic policy consists of π_1^* and π_2^{**}, compared with the time-consistent policy path of π_1^* and π_2^*, so that both policy paths share

the same policy in period 1 but have different policies in period 2. The difference between π_2^* and π_2^{**} arises because of the difference between (23) and (17). (23) will differ from (17) unless:

$$\frac{\partial x_1}{\partial \pi_2}\left[\frac{\partial U}{\partial x_1}+\frac{\partial U}{\partial x_2}\cdot\frac{\partial x_2}{\partial x_1}\right]=0 \tag{24}$$

If (24) is not satisfied, which it will not be for many forms of the $U(.)$ and $x_2(.)$ functions, $\pi^{**}{}_2$ will differ from $\pi^*{}_2$. Since the consistent policy path $(\pi^*{}_1, \pi^*{}_2)$ maximizes intertemporal utility, $(\pi^*{}_1, \pi^{**}{}_2)$ will yield a lower utility level from the perspective of both periods 1 and 2 since the utility function is the same for both periods. That is, $U(x_1, x_2, \pi^*{}_1, \pi^*{}_2) \leq U = U(x_1, x_2, \pi^*{}_1, \pi^{**}{}_2)$. Hence the discretionary myopic policy path is inferior to the time-consistent one. As explained earlier, the reason for this suboptimality is the passage of time, and that the discretionary policies adopted in period 2 ignore the impact of such policies, had they been known in advance, on *past* decisions on π_1 and x_1. Hence, *ex ante* maximization of the policy maker's utility function requires that it must pursue $\pi_2{}^*$ in period 2, even though reoptimization (but with one period less in the problem) at the beginning of period 2 would yield a different policy π_2^{**}.

In practice, economic agents might know $\pi_2{}^*$ and $\pi_2{}^{**}$, and some would find it to their advantage to have the latter rather than the former pursued, thereby leading to pressure on the policy maker to pursue $\pi_2{}^{**}$. However, *ex ante* intertemporal optimization suggests that the policy maker must resist such pressure. If the policy maker has pre-announced its future policies, it must resist public pressure to change them and should maintain a credible reputation for sticking to any pre-announced policies.

Impact of a rolling horizon on the optimal policy

In the context of a fixed two-period horizon, optimization in the second period becomes myopic, which is a factor in the derivation of the Kydland and Prescott (1977) results. However, there is no reason to assume that the central bank will switch from a two-period to a one-period horizon (and then none!) over time. It is more likely that the length of the horizon will remain constant, so that the appropriate analysis will be a rolling one. In the two-period context, maintaining the horizon at two periods means a rolling horizon of two periods, so that reoptimization at the beginning of the second period would also be over two periods. Consequently, if there is no change in the objective function and its constraints, there will also be no change in the policies derived under optimization at the beginning of period 1 and of period 2 – as well as at the beginning of periods 3, 4 and so on.

A similar argument can be made for longer rolling horizons, as we discuss later.

12.4.3 Limitations on the superiority of time-consistent policies over reoptimization policies

Note that the preceding two-period analysis assumed that the policy maker's preference function did not shift over time nor was there a shift in the economy's response functions, so that the demonstrated superiority of time-consistent over discretionary policies was not due to

shifts in preferences and/or in private agents' response functions. These are the fundamentals of the analysis. If these do shift, the optimal policy path will change, without any implication about the merits of one policy over another.

Three considerations on the practical relevance of the superiority of time-consistent over discretionary policies need to be noted. One of these is related to the extent of the deviation of $\pi^{**}{}_2$ from $\pi^{*}{}_2$. The following discussion provides some intuition on this deviation. If there are only two periods, as in the preceding analysis, the transition from period 1 to 2 affects about 50 percent of the information, so that there could be a very significant difference between $\pi^{**}{}_2$ and $\pi^{*}{}_2$.[5] But if the horizon is ten years away, the passage of one year will affect only about 10 percent of the information. In a rolling system, with the horizon always being ten years and with a given unchanging structure of the economy, at the end of the first year that year will be replaced by a similar tenth year, so that the difference between $\pi^{**}{}_2$ and $\pi^{*}{}_2$ could be even smaller. Knowledge in economics is never precise enough to allow a high degree of precision, so that a small shift in the optimal policies need not have much practical significance in terms of the policies pursued or their impact; thus the difference between time-consistent and reoptimization policies may be of little practical significance.

The second consideration is that, under uncertainty, a time-consistent policy is likely to prescribe for future periods complicated policies that depend on the outcomes in earlier periods. The central bank could avoid such complicated policy prescriptions by merely announcing a commitment to its intertemporal utility function and intertemporal reoptimization each period, but not announce the interest rates and money supply it would engineer for each period. This would avoid a complex pre-specification of future policies, as well as avoid having to stick rigidly to these policies if circumstances change in a way that makes them inappropriate.

Third, given the existence of uncertainty, the knowledge of the economy and its future evolution do tend to change with the passage of time. A mere change in knowledge of the economy would change the constraints in the optimization exercise by changing the perceived likelihood of outcomes and/or their probabilities. Further, unanticipated shifts in the economy – as well as unanticipated shifts in foreign economies that impact on it – occur over time. In the preceding model, the unanticipated shifts would be represented by a shift in $x_1(\pi_1, \pi_2)$ and/or in $x_2(x_1, \pi_1, \pi_2)$. Let this shift be such that, under the discretionary procedure, the reoptimization in period 2 implies the optimal policy $x^{***}{}_2$, different from both $x^{*}{}_2$ and $x^{**}{}_2$.[6] The level of utility attained with $x^{*}{}_2$ and $x^{***}{}_2$ could well be higher or lower than with $x^{*}{}_2$ and $x^{**}{}_2$, so that no general judgment can be rendered as to which policy path is superior.[7]

Hence, the above analysis of the superiority of time-consistent policies does not strictly apply if knowledge of the actual constraints or shifts in the constraints themselves occur, so that a policy maker who sticks rigidly to the a priori time-consistent policy path determined in the past could end up pursuing inappropriate policies. The greater the change in information

5 In the two-period analysis, optimization in the second period becomes a myopic, rather than an intertemporal, one, which drops the subjective discount and interest rate variables out of the optimization. This would not occur under a long horizon or a rolling one.

6 $\pi^{**}{}_2$ differs from $\pi^{*}{}_2$ because of the information shift with the passage of time, while $\pi^{***}{}_2$ differs from $\pi^{**}{}_2$ because of the shift in the economy's response functions $x_1(.)$ and $x_2(.)$.

7 In addition, for the new $x_2(.)$ constraint, $\pi^{*}{}_2$ and $\pi^{**}{}_2$ would no longer represent the optimal policy path.

and the more significant the unanticipated shocks[8] to the economy, the less relevant would be the practical importance of the analytical superiority of time-consistent policies. Intuition seems to indicate that reoptimization policies, even with an unchanging intertemporal utility function, are likely to become superior if changes in the knowledge of the economy and/or unanticipated shocks are significant.[9]

Knowledge of the future course of the economy is virtually always inadequate and new information does arrive in every period, so that, factoring in the new information, reoptimization will almost always tend to yield greater welfare than continuing with the time-consistent policy path derived some periods earlier. Given the impossibility of predicting the probabilities of future outcomes accurately, reoptimization becomes the superior procedure by leaving a scope for flexibility and judgment in policies while maintaining the same objective function over time.

To conclude, while the analytical superiority of time-consistent over discretionary policies can be shown under certain assumptions, their practical superiority cannot be taken for granted. In fact, in cases of limited, vague and imperfect information on future periods and significant unanticipated shocks, it seems likely that the reoptimization policy procedure could prove to be superior. Further, even if there were no new information, a long rolling horizon would yield policies that are more or less the same as under once-for-all optimization.

Empirical relevance of intertemporal optimization procedures

The empirical relevance of time consistent policies is limited in various ways. For one, it is hard to find a central bank that gives a commitment to maintaining a time-consistent policy path. Two, as argued above, there are no gains from such a policy or, if there are any, they are of a second or third degree of significance relative to a policy of reoptimization with a maintained objective function under a longish rolling horizon, which is what many central banks seem to follow. Three, under shifts in the economy and/or information about it, continual reoptimization can provide greater welfare than time consistency. Four, some economists have argued that if there were serious losses from failure to follow time-consistent policies with a commitment to zero inflation, central banks, governments and society would have realized it and would have implemented procedures to follow such policies, thereby avoiding the implied losses.

Therefore, in practice, many central banks seem to follow intertemporal reoptimization procedures with adherence to unchanging objective functions rather than to time-consistent policies.

Using the Taylor rule to illustrate time-consistent versus reoptimization policies

Suppose that the central bank has the twin goals of reducing the output gap and the deviation of inflation from its desired level ("inflation gap") and maximizes its specific intertemporal objective function (see next section) subject to the economy's constraints over a long and rolling horizon. This would imply a specific time-consistent policy path for the current and

8 Unanticipated shocks are ones that shift the outcomes and/or their probabilities or/and the economy's response functions, other than those that occur due to the mere passage of time.

9 That is, the present discounted value of utility attained under a fixed time-consistent policy path would be less than that under a discretionary policy that had taken account of the changes in information and the shocks.

future periods. Suppose that this time-consistent path is that specified by the Taylor rule with *specific* fixed weights on the output gap and the inflation gap.

Now suppose that the central bank maintains an unchanged objective function but reoptimizes each period over the long and rolling horizon. If economic conditions do not change, its optimal policy path, as specified by the Taylor rule with fixed coefficients, would remain identical to the time-consistent path. But if the economic conditions or their perception do change, then the constraints will shift. Reoptimization is then likely to imply that, while the general form of the Taylor rule will not change, its coefficients will. In particular, the weight on the output gap relative to that on the inflation gap will differ from that in the time-consistent Taylor rule. That is, reoptimization allows the central banker to change these relative weights as economic conditions or their perception shift. A shift of this nature would also occur if the goal variables do not change but the parameters of the objective function shift due to a change in the leadership of the central bank.

An illustration: time consistency versus reoptimization for monetary policy

Assume that the objectives of the central bank include output in periods 1 and 2 because this would promote output growth under a two-period horizon. Further, assume that output depends positively on investment with a one-period gestation lag, and investment is negatively related to inflation above a 2 percent rate. Now suppose that the central bank's welfare function over the two periods is maximized if the central bank gives a credible commitment that inflation will be maintained in both periods at 2 percent, so that the public's expected inflation rate for period 2 is 2 percent. That is, the resulting levels of investment and output under the monetary policy of an inflation rate of 2 percent maintained over the two periods maximize the central bank's discounted level of utility over the two periods. Consequently, the central bank engineers inflation at 2 percent in period 1. Under a time-consistent policy, it would pursue a monetary policy in period 2 that also produces a 2 percent inflation rate in period 2.

Now suppose that the central bank were to follow a reoptimization policy procedure and, at the beginning of period 2, would reoptimize its utility function. Under a fixed two-period horizon that started in period 1, this reoptimization would be only over period 2, since period 1 would now be in the past. Since period 2's investment does not affect period 2's output because of the gestation lag, the central bank feels free to change its monetary policy and the resulting inflation from the 2 percent rate. Therefore, in period 2, without affecting its output, the central bank can accede to pressure from the government facing an election, or for financing the fiscal deficit, and abandon its policy of maintaining the 2 percent inflation rate. Let this imply a monetary policy in period 2 that causes the inflation rate to become 10 percent. If the central bank's objective function includes *only* investment and its effect on output in periods 1 and 2, the fall in investment in period 2 will reduce its attained utility level, so that the deviation from the time-consistent policy path of 2 percent inflation will be inferior (i.e. yield less utility) to the 2 percent inflation path. In addition, there is a loss in credibility, which would have implications beyond period 2. Therefore, sticking to the time-consistent policy of 2 percent inflation each period would be preferable.

Now consider the extension of this analysis to reoptimization under a rolling two-period horizon with unchanged objective function and constraints. Reoptimization in period 2 would yield the same optimal policy path as that at the beginning of period 1. Hence, the policies pursued under reoptimization in each period would be identical with those under

a once-for-all optimization. This conclusion would also hold if the number of periods were increased beyond two. Hence, discretion in the sense of reoptimization will not yield lower utility than time-consistent policies

However, reoptimization will allow the flexibility of modifying the policy path if the objective function and/or the perception of the constraints shift. An illustration of such shifts occurred in late 2007 with the onset of the subprime financial crisis in the USA. Information on its extent and severity kept changing weekly, if not daily. In response to this crisis, the Fed reacted by reversing its earlier policy path, which had been one of monetary tightening through increases in the interest rate. As the perception of the crisis in the mortgage and financial markets became more dismal, the Fed resorted again and again to lowering the interest rate and pumping liquidity into the economy. There was no precedent in the Fed's policies for such actions, and a time-consistent policy, formulated as late as early 2007, would have been against it.

12.4.4 Inflationary bias of myopic optimization versus intertemporal optimization

Intertemporal optimization of either the time-consistent type or period-by-period reoptimization can also be compared with myopic (one period, short-run) optimization subject to the current one period (short-run) constraint. As a result of the time consistency debate, the inferiority of the latter is now widely accepted. It has also been argued above that such a policy can have an inflationary bias if the objective is to achieve a level of output above the full-employment level when the economy does not allow such a possibility.

To illustrate, if we assume that the constraint is of the original Phillips curve type (see Chapter 15), with a negative tradeoff between inflation and unemployment, optimization of the central bank's utility function over inflation and unemployment is likely to imply the choice of a positive rate of inflation, say $\pi_1{}^*$, and an unemployment rate $u_1{}^*$ above the natural rate u^{n}. However, if we now assume that there is no long-run tradeoff between anticipated inflation and unemployment, that rational expectations hold and that the short-run tradeoff is really the expectations-augmented Phillips curve (see Chapter 14), attempts to maintain unemployment at $u_1{}^*$ will mean that, over time, the central bank will have to pursue increasingly more expansionary policies, resulting in accelerating inflation rates. Policy pursued in the posited manner has an *inflationary bias*, without a long-run reduction in unemployment. This bias arises because of the possibility that the central bank, optimizing over one period and subject to the short-run Phillips curve, chooses an unemployment rate above the natural one when, in fact, the economy does not allow this possibility over time. This inflationary bias really arises from the central bank's error in assuming that the economy follows a simple Phillips curve when in fact it pursues an expectations-augmented one.

An inflationary bias is unlikely to occur under the two intertemporal optimization procedures with a long or rolling horizon since the latter would embody the impact of current inflation on the inflation expected for future periods, so that the losses from future inflation will be captured in the current intertemporal utility function.

12.4.5 Time consistency debate: modern classical versus Keynesian approaches

The above analyses show that the differences between the Keynesian and the classical paradigms rest on several distinct elements: short-term optimization versus intertemporal

optimization, time consistency (stationary optimization) versus reoptimization policies, the nature (subjective or objective) of probabilities and its impact on revisions in information, the possibility of unanticipated shocks, and multiple one-period constraints versus a long-run one. Of these, the clear achievement of the time consistency debate has been to effectively eliminate arbitrary and one-period myopic optimization subject to the short-run Phillips curve, which had been advocated by the Keynesians in the 1950s and 1960s, because it is not intertemporally optimal and has an inflationary bias. Since the economy is a continuing entity, with consequences of the present on the future, goals and policies determined under the limitation of a one-period horizon, are neither realistic in terms of how most central banks usually behave and how economies function, nor are they now recommended by the new Keynesians or any other major school.

Further, few economists, classical or Keynesian, advocate arbitrary shifts over time in the central bank's objective function. Their disputes revolve around the economy's performance and the tradeoffs that the economy permits or does not permit over time, and the potential for future shifts in the economy. On the last point, modern classical economists usually view the policy maker's probabilities of future outcomes as fairly accurate and therefore close to the objective estimates. Given such prior beliefs, they support time-consistent policies as being the superior ones for central banks to follow. By comparison, the Keynesian paradigm has historically taken the view that the central bank's information on future outcomes is vague and imperfect, so that the subjective probabilities it holds can be quite different from objective probabilities and can shift as new information becomes available. Further, the central bank has knowledge of these limitations (for example, it does not know what it does not know) and admits the possibility of errors in its subjective probabilities (without knowledge of what the errors will be and how to correct for them *ex ante*) and of potential future revisions in them. Hence, the Keynesians recommend giving the central bank a free hand to reformulate its policies anew each period on the basis of intertemporal reoptimization. Note, however, that the Keynesians do not espouse shifts in the goal function of the policy maker, so that changes in the policies pursued results not from a shift in its preferences over time but from newer and better information emerging on the future state of the economy. This becomes a recommendation for reoptimization each period so as to capture the continual improvements in information. Reoptimization, therefore, has a strong economic rationale whenever information is likely to change significantly over time.

Therefore, the general Keynesian justification for giving the central bank discretion to continually reformulate, if needed, its policies on the money supply and interest rates (though not its ultimate objective function) rests on a foundation not addressed in the proof of the superiority of time-consistent over reoptimization policies. As shown earlier, with unanticipated changes in information occurring over time, the Keynesian reoptimization procedure is likely to yield superior policies and is the one currently being followed by most central banks.

12.4.6 Objective functions for the central bank and the economy's constraints

Objective functions

The objective functions considered in this section assume that the central bank aims at stabilization of the inflation rate rather than the price level. As discussed in Chapter 10, under a price level target the central bank will have to offset any deviations from this target,

so that a price increase in one period will require the central bank to follow deflationary policies in a subsequent period. Under an inflation objective the price level increase will not require such deflation. The former policy is therefore likely to induce greater volatility of output and inflation than the latter, though it will also produce greater stability of the price level over time, which will lower long-term price uncertainty. The central banks of most countries, including Canada, the UK and the USA, choose to target inflation rather than the price level.

The objective function $U(.)$ can be specified as linear or non-linear. Assuming that the central bank is only concerned with deviations of output and inflation from their desired levels, a mixed linear and quadratic form of the objective function is:

$$U = \lambda\{y - (y^{\mathrm{f}} + k)\} - \tfrac{1}{2}(\pi - \pi^{\mathrm{T}})^2 \qquad k \geq 0, \lambda > 0 \tag{25}$$

Under this objective function, the central bank attaches positive marginal utility to output but negative utility to the inflation gap. This objective function is linear in the "*output gap*" $(y - (y^{\mathrm{f}} + k))$, where the desired level of output is $(y^{\mathrm{f}} + k)$, but quadratic in the deviation of inflation from its target level π^*. For convenience, $(\pi - \pi^{\mathrm{T}})$ will be called the "*inflation gap*." λ is the relative weight placed on the output gap versus the inflation gap. A lower value of λ means that the weight attached to output increases would be less than to increases in inflation, so that a central bank with a lower value of λ would follow a more stringent anti-inflationary policy than one with a higher value. In this context, the literature defines a "*conservative banker*" as one with a lower value of λ than that assigned by the public.

Since (25) attaches positive utility to the output gap rather than to its quadratic value, it implies that the central bank does not seek to *stabilize* output at y^{f} or $(y^{\mathrm{f}} + k)$. If $k > 0$, its output target is above y^{f}, which may be justified by the reasoning that y^{f} is lower than it should be because of labor market rigidities, monopolistic competition and bottlenecks in the economy. A positive value can also arise from public and governmental pressures for a lower unemployment than the economy can deliver at the full-employment output. Such pressure could arise because the government believes that the unemployment rate is unacceptably high and/or reducing it will reduce poverty, or because its re-election chances might be hurt by high unemployment.

An alternative form of the objective function in more common use is quadratic in both terms, as in:

$$U = -\tfrac{1}{2}\lambda\{y - (y^{\mathrm{f}} + k)\}^2 - \tfrac{1}{2}(\pi - \pi^{\mathrm{T}})^2 \tag{26}$$[10]

Whereas (25) included only a role for inflation stabilization around its desired level, (26) also introduces a role for output stabilization around $(y^{\mathrm{f}} + k)$, since any such deviations involve a loss of utility for the central bank. If $k = 0$, the central bank wants to stabilize output around its full-employment level. But if $k > 0$, the central bank wants to stabilize output at $(y^{\mathrm{f}} + k)$. The interesting question that arises in this context is whether the constraints imposed by the economy permit it to do so. The usual long-run constraint does not allow it, but short-run constraints usually do.

10 Expanding this function yields $U = \lambda k\{y - y^{\mathrm{f}}) - \tfrac{1}{2}(\pi - \pi^*)^2 - \tfrac{1}{2}\lambda\{y - y^{\mathrm{f}})^2 - \tfrac{1}{2}k^2$, where the first term allows for positive utility for output greater than the full-employment level, while the third term introduces a loss from deviations of output from the full-employment level and thereby introduces a role for output stabilization policies.

An illustration of the quadratic intertemporal objective function is provided by the new Keynesian function (see Chapter 15):

$$-\tfrac{1}{2}E_t\left\{\sum_{j=0}^{\infty}\beta^j\left(\lambda(x_{t+j}^{\#}-k)^2+\pi_{t+j}^{\#2}\right)\right\} \qquad k\geq 0 \tag{27}$$

where $x^{\#}$ is the output gap ($=y-y^{\mathrm{f}}$), so that the target output level is again taken to be the full-employment level plus k (i.e. $x^{\#}=y-(y^{\mathrm{f}}+k)$), and $\pi^{\#}$ is the inflation gap, defined as the deviation of inflation from its desired target level π^{T}.[11] β is the central bank's time discount factor and λ is the weight placed by the central bank on the output gap relative to the weight on the inflation gap. This objective function is maximized subject to the economy's long-run constraint or the set of short-run constraints for the current and future periods.

The preceding objective functions are relevant if the central bank is concerned only with output and inflation. However, a central bank may also wish to include other variables in its objective function. Among these can be exchange rate variability, which could be of concern for a highly open economy with floating exchange rates (Ball, 1999), and money supply variability, which could make the financial markets more volatile. These would require making suitable modifications to the central bank's objective function.

Usual specification of the economy's supply side constraints

Optimization requires that the objective function be maximized subject to the relevant constraints imposed by the economy on the supply relationship between output and inflation. There are both long-run and short-run versions of these constraints. The usual long-run supply constraint specifies that $y^{\mathrm{LR}}=y^{\mathrm{f}}$, which states that the economy's long-run output cannot differ from the full-employment one. Both the modern classical and Keynesians subscribe to this long-run constraint. However, differences among them arise on the appropriate form of the short-run supply relationship. For the 1950s and 1960s Keynesians it was the Phillips curve, which has the form:

$$u=f(\pi) \qquad \partial u/\partial\pi<0$$

where u is the unemployment rate. Since unemployment and output are negatively related, the PC can also be written as the aggregate supply equation:

$$y=f(\pi) \qquad \partial y/\partial\pi>0$$

While the original Phillips curve drawn in the (u, π) space is negatively sloped, that between y and π in the (y, π) space will be positively sloped. A linear or log-linear form of this

11 Some studies assume that the central bank's objective function is identical with the welfare function of the public. In such a function, used by Rotemberg and Woodford (1999), the deviation of inflation from its trend (rather than a pre-set desired inflation rate) has negative utility because this deviation makes it more difficult for economic agents to plan for consumption, investment and portfolio allocations. The individual's welfare also is a function of the output gap since this gap is positively related to fluctuations in jobs and incomes.

version of the Phillips curve is:

$$y = y^{\mathrm{f}} + \alpha\pi \qquad \alpha > 0$$

The Phillips curve was a relationship between the actual values of π, u and y.

For the modern classical school, the appropriate form of the supply constraint is the Friedman–Lucas one (see Chapters 14 and 16), whose linear or log-linear form is:

$$y^{*} = y^{\mathrm{f}} + \alpha(\pi - \pi^{\mathrm{e}}) \qquad \alpha > 0 \tag{28}$$

where * indicates the short-run equilibrium level, assumed by classical economists to be the actual level, in competitive markets. Under rational expectations, π^{e} is replaced by its rationally expected value $E\pi$. This short-run aggregate supply function is also known as the expectations-augmented Phillips curve (EAPC), which occurs because workers sign contracts for nominal wages on the basis of the price level expected during the contract period, or because firms misinterpret an increase in their prices as including a relative price increase when it actually does not do so (see Chapter 14). In the long run (by virtue of its definition that it is the state in which there are no errors in expectations), $\pi^{\mathrm{e}} = \pi$, so that the long-run supply constraint is simply:

$$y^{\mathrm{LR}} = y^{\mathrm{f}}$$

While the Friedman–Lucas supply equation is the usual specification of the aggregate supply in time consistency and credibility analyses, the general assessment in the literature (Lucas, 1996) is that, in practice, "only small fractions" of the actual deviations of output from its full-employment level occur due to errors in inflationary expectations, so that (28) is not a satisfactory equation for explaining (i) the departures from full-employment output, (ii) the way in which monetary policy impacts on output, or (iii) the extent of that impact.

The Friedman–Lucas aggregate supply function is generally not acceptable to Keynesians and new Keynesians (see Chapter 15), who allow for the impact of demand increases on output without a prior impact on inflation. The new Keynesian (NK) models propose a price-cum-output adjustment constraint derived from intertemporal behavior of the economy with monopolistic firms and staggered price setting (see Chapter 15). One form of such a constraint (sometimes referred to as the new Keynesian Phillips curve) is specified (Clarida *et al*., 1999) as:

$$\pi_t = \alpha(y_t - y^{\mathrm{f}}) + \beta\pi^{\mathrm{e}}_{t+1} + z_t \qquad \alpha, \beta > 0 \tag{29}$$

where z represents supply price shocks, such as those from increases in prices of inputs (including labor) to inflation and represents the contribution of cost-push inflation. A positive output shock that increases the full-employment output decreases inflation through the output gap. The NK supply function can also be rewritten as:

$$y_t = y^{\mathrm{f}} + \frac{1}{\alpha}\pi_t - \frac{\beta}{\alpha}\pi^{\mathrm{e}}{}_{t+1} + \frac{1}{\alpha}z_t \qquad \alpha, \beta > 0 \tag{30}$$

Under the rational expectations hypothesis, $\pi^{\mathrm{e}}{}_{t+1} = E_t\pi_{t+1}$. Taking period t as the current period, note that, in the NK equation, current output is a positive function of current inflation

but a negative function of expected future inflation, not of expected inflation in the current period t. New Keynesians accept that for the long run $y = y^{f}$ and $\pi = \pi^{e}$, so that, from (29), while inflation can occur in the long run, output will be invariant with respect to the inflation rate.

Economy's demand side constraints

If the central bank keeps the money supply constant on an exogenous basis, the relevant demand side constraints are provided by the IS and LM equations in the case of monetary targeting and by the IS equation in the case of interest rate targeting.

For the determination of the optimal money supply under monetary targeting, some models avoid the complexity of the IS–LM model by assuming a direct quantity theory link between inflation and the money supply, as specified by:

$$\pi = \Delta M + \eta$$

where ΔM is the growth rate of the nominal money supply and η is a disturbance term. However, such an assumption seems inappropriate for short-run analyses that allow the impact of money supply changes on real output. It is also inappropriate if the money demand function is unstable and the central bank uses the interest rate as its primary monetary policy instrument. If the central bank targets the market interest rate, the relevant demand side constraints are provided by the IS equation and the exogenously specified interest rate (see Chapter 13).

We bypass the issue of whether the money supply or the interest rate is held constant exogenously by assuming that the central bank controls aggregate demand on an exogenous basis. In this case, one specification of the link between inflation and aggregate demand would be:

$$\pi = \Delta y^{d} + \eta \tag{31}$$

where Δy^{d} is the growth rate of aggregate demand at the existing price level and η is the error in the central bank's control of inflation through its manipulation of aggregate demand. Note that this is quite a limiting assumption, which is also not consistent with Keynesian ideas. A more realistic relationship relevant to the short run in which output can respond to aggregate demand changes would be:

$$\pi = b\,\Delta y^{d} + \eta \qquad 0 \leq b \leq 1 \tag{32}$$

In terms of realism, a demand-deficient recession, with output below its full-employment level, would have $b < 1$. Further, the usual experience of even business cycle upturns fuelled by aggregate demand increases is that output increases for some time prior to increases in inflation (Mankiw, 2001), so that if the short-run models are to explain this occurrence, b must be less than unity for some time and adjust upwards over time. This formulation of the link between aggregate demand and inflation also allows the possibility that not all demand increases impact on output through price or inflation increases (Lucas, 1996; also see Chapter 14, Conclusions). However, this possibility makes the Friedman–Lucas supply constraint inapplicable, even though it represents the usual depiction of the economy's supply function in the time consistency debate.

The appendix to this chapter provides some mathematics on both myopic and intertemporal optimization for the optimization problem set out in this section. We also use below the objective functions and the expectations-augmented Phillips curve to analyze the inflationary bias of discretionary policies and the utility loss from them, and of credibility.

12.5 Commitment and credibility of monetary policy

The discussion above of the time consistency debate touched on the related but distinct concept of credibility, which is also very relevant in the economy's response to the central bank's policies. To illustrate this concept for monetary policy, start with the central bank's announcement in period 1 that it would maintain a 2 percent inflation rate. Now suppose that the central bank took the inflation rate to 10 percent through an expansionary monetary policy in period 2. Let the adverse effects of this rate on growth induce the central bank to announce and implement the reduction of the inflation rate in period 3 back to 2 percent. But the increase in the inflation rate in period 2 would have caused the public to doubt the central bank's commitment to the 2 percent inflation objective and disdain any further rhetoric about it, so that it would continue to expect higher inflation rates for period 3. As a result, investment and growth are likely to be lower and inflation higher for several periods than if the central bank had stayed with its original commitment. This is essentially an argument for maintaining the credibility of policy. This argument implies that the central bank should resist its own temptation, as well as pressures from various groups, including the government, to vary its inflation target. Further, a high level of credibility requires that the central bank should maintain the transparency of its policies, for example by announcing clearly specified policies for the future and sticking to them. Changes in the announced policies over time, especially under conditions that were expected, would be suboptimal.

The reputation and credibility of the central bank is an important topic in policy analysis (Barro and Gordon, 1983; Fischer, 1990) since economic agents take their perception of the central bank's actions into account in forming their expectations and determining wage negotiations, consumption, investment, production, etc. The bank's credibility can be established through an acquired reputation of sticking to its pre-announced policies or objectives. At the level of objectives, the credibility of the central bank's announced (or perceived) objectives depends on its commitment to them and its ability to achieve them. This credibility is easier to establish if the central bank has only one objective, such as an inflation target, rather than independent or conflicting multiple objectives, since their established target levels are likely to be more difficult, if not impossible, to achieve simultaneously.[12] At the level of policies, a credible policy announced by the central bank is one in which the public has faith because it expects the policy to be implemented and believes that it will attain its declared objectives.

12.5.1 Expectations, credibility and the loss from discretion versus commitment

To illustrate the inflationary bias and loss in utility from discretion (under which the central bank does not commit itself to following an aggregate demand policy) versus commitment

12 While the Taylor rule has only two, interrelated, targets (the inflation rate and full-employment output), it often becomes difficult to achieve both, so that a tradeoff has to be exercised between inflation and the output level. Favoring the latter level may mean loss of credibility on fighting inflation.

(where it commits itself to following a policy that ensures a zero inflation rate on average), we present two alternative analyses of this issue with slight variations in the objective function and the budget constraint.

For the first analysis, assume that the central bank has the objective function:

$$U = \lambda(y - y^{\mathrm{f}}) - \tfrac{1}{2}\pi^2 \qquad \lambda > 0 \tag{33}$$

and the economy imposes the Friedman–Lucas supply constraint:

$$y = y^{\mathrm{f}} + \alpha(\pi - \pi^{\mathrm{e}}) + \mu \qquad \alpha > 0 \tag{34}$$

Note that the objective function is linear in the output gap, the target output is y^f and the target inflation rate is zero, and that the constraint includes a disturbance term μ. The economy's constraint from the demand side is specified as:

$$\pi = \Delta y^{\mathrm{d}} + \eta \tag{35}$$

where μ and η are random errors, with a zero mean, and Δy^{d} is the growth rate of aggregate demand. All variables are in logs. Assume that the public can observe the central bank policy of ensuring Δy^{d} but are not able to observe η prior to forming their expectations (as in the Friedman and Lucas supply analyses, see Chapters 14 and 17). Substituting (34) and (35) in (33) yields:

$$U = \lambda[\alpha(\Delta y^{\mathrm{d}} + \eta - \pi^{\mathrm{e}}) + \mu] - \tfrac{1}{2}(\Delta y^{\mathrm{d}} + \eta)^2 \qquad \lambda, \alpha > 0 \tag{36}$$

Taking π^{e} as given, maximization with respect to y^{d} yields:

$$\Delta y^{\mathrm{d}} = \lambda\alpha \qquad \lambda, \alpha > 0 \tag{37}$$

which, inserted in (36), yields π as:

$$\pi = \lambda\alpha + \eta \qquad \lambda, \alpha > 0 \tag{38}$$

Since the public was assumed to know the central bank policy on aggregate demand but not to know η prior to forming their expectations, the public's expected inflation rate under rational expectations is:

$$\pi^{\mathrm{e}} = E(\pi) = \Delta y^{\mathrm{d}} = \lambda\alpha \qquad \lambda, \alpha > 0 \tag{39}$$

Hence, the expected inflation rate is $\lambda\alpha$. From (34), (38) and (39), we have:

$$y_t = y^{\mathrm{f}} + \alpha\eta_t + \mu_t \qquad \lambda, \alpha > 0 \tag{40}$$

In (40), since $Ey_t = y^{\mathrm{f}}$, there is no benefit *on average* in terms of output. However, even without such a benefit, (38) shows that the central bank will inflate aggregate demand and cause inflation on average at the rate $\lambda\alpha$. The larger α is (the gains from unanticipated

inflation due to its impact on output), the greater is this inflation. Under rational expectations, the public recognizes this tendency and will anticipate a higher rate of inflation for higher values of α.

For the preceding problem and its optimal policy, the expected utility of the central bank is given by:

$$\begin{aligned} E[U^{\mathrm{d}}] &= E[\lambda(\alpha\eta + \mu) - \tfrac{1}{2}(\lambda\alpha + \eta)^2] \\ &= -\tfrac{1}{2}[\lambda^2\alpha^2 + \sigma^2{}_\eta] \end{aligned} \tag{41}$$

where $\sigma^2{}_\eta$ is the variance of the error η in the central bank's control of inflation and the superscript d on U indicates the utility level under discretion.

We want to compare the above inflation rate and utility level attained under discretion with the case where, given that the full-employment output has a zero growth rate, the central bank commits itself to keeping demand growth equal to zero, i.e. $\Delta y^{\mathrm{d}} = 0$, so as to produce $E(\pi) = 0$. In this case of commitment to a zero inflation rate on average, the actual inflation rate π is:

$$\pi^c = \eta \tag{42}$$

where the superscript c stands for commitment. Comparing the attained inflation rates under discretion and commitment, inflation under the latter is systematically less by $\alpha\lambda$.

The expected utility attained under commitment, $E[U^c]$, is given by:

$$\begin{aligned} E[U^{\mathrm{c}}] &= E[\lambda(\alpha\eta + \mu) - \tfrac{1}{2}\eta^2] \\ &= -\tfrac{1}{2}\sigma^2{}_\eta \end{aligned} \tag{43}$$

Since $\lambda^2\alpha^2 > 0$, $E[U^{\mathrm{c}}] > E[U^{\mathrm{d}}]$, there is a gain in utility from commitment. Hence, discretion leads on average to higher inflation (the "inflation bias" of discretionary policy) as well as to lower utility for the central bank – and for society if the two have the same utility function.[13]

An alternative analysis

The following presents a slightly different analysis with an objective function quadratic in both the output gap and in the deviation of inflation from its target value, and without a disturbance term in the supply constraint. The demand constraint has also been omitted. For this illustration, assume that the central bank maximizes the quadratic objective/utility function:

$$U(\pi_t, y_t) = -(\pi_t - \pi^{\mathrm{T}})^2 - \lambda(y_t - y^{\mathrm{f}})^2 \tag{44}$$

13 This result also holds when the utility function includes quadratic terms in the deviations of both output and inflation from their desired levels.

where π^{T} is the long-run desired inflation rate.[14] Assume, as above, that the central bank's perceived – and the economy's actual – aggregate supply relationship is of the Friedman–Lucas (expectations-augmented Phillips curve) type:[15]

$$y_t = y^{\mathrm{f}} + \alpha(\pi_t - \pi^{\mathrm{e}}{}_t) \qquad \alpha > 0 \tag{45}$$

For myopic optimization, the expected inflation rate is taken to be given (i.e. it is independent of π_t and y_t) and the central bank optimizes with respect to π_t and y_t, taking these given values of $\pi^{\mathrm{e}}{}_{\mathrm{t}}$. π^{T} and y^{f} to be exogenous to this model. Substituting for y_t from (45) in (44) yields:

$$U(\pi_t, y_t) = -(\pi_t - \pi^{\mathrm{T}})^2 - \lambda(\alpha(\pi_t - \pi^{\mathrm{e}}{}_t))^2 \tag{46}$$

Maximizing this function with respect to π_t yields:

$$\pi_t = (\pi^{\mathrm{T}}{}_t + \alpha\lambda\pi^{\mathrm{e}}{}_t)/(1 + \alpha\lambda) \tag{47}$$

Substituting this value of π_t, for given π^{e}, in the objective function (46) yields negative utility (welfare loss).

If the central bank had followed long-run rather than myopic optimization, it would have had to treat the expected inflation rate as endogenous. If we assumed the full credibility of the central bank's policy of maintaining inflation at its desired rate, $\pi^{\mathrm{e}}{}_t$ would equal $\pi^{\mathrm{T}}{}_t$. Inserting this in equation (47) yields the long-run result that $\pi_t = \pi^{\mathrm{T}}{}_t$. Substituting this result in the supply function yields y equal to y^{f}. Therefore, the objective function (46) yields a zero welfare loss. Hence, the failure of the central bank to give an advance commitment to achieving its target inflation rate π^* for all future periods results in a higher welfare loss than a credible advance commitment. The bank's mistake in following a myopic policy is that it fails to realize that it can shape expectations and the Phillips curve tradeoff between inflation and output by providing a prior commitment to its policy.

Imposition of goals, constraints and incentives on central banks

The discussion so far in this chapter has shown two distinct sources of inflation bias.[16] These are:

- Policy pursued under an incorrect version of the economy's supply and demand constraint or constraints. In this category would fall the use of the Phillips curve, as against the

14 Depending on the structure of the economy, π^T is replaced in some such equations by the last period's inflation rate.

15 Woodford (2007, equation 2.1) offers the following intertemporal version of this constraint as the new Keynesian aggregate supply function:

$$\pi_t - \pi^*{}_t = \alpha \ln(y_t / y^{\mathrm{f}}{}_t) + \beta E_t(\pi_{t+1} - \pi^*{}_{t+1}) + \mu_t \qquad \alpha > 0, 0 < \beta < 1$$

where $\pi^*{}_t$ is the perceived trend of inflation rate in period t.

16 For the open economy, the central bank may choose to have an objective function that has been modified to include disutility of a deviation of the exchange rate (or the balance of payments balance) from its desired level, or its variance. With domestic inflation causing depreciation of the exchange rate, the additional term will tend to reduce the inflationary bias of discretionary policy.

use of the expectations-augmented Phillips curve. This inflation bias would be due to an error in the central bank's knowledge of the economy.

- Policy pursued under the expectations-augmented Phillips curve when the central bank does not commit itself to ensuring aggregate demand that will maintain zero inflation on average and/or does not have full credibility.

The second source of an inflation bias leads to issues of how to induce or ensure that the central bank will follow a zero inflation policy. The literature investigates several scenarios for achieving it. The simplest would be to mandate the central bank to follow such a policy.[17] Another possibility is to provide the central banker with a positive incentive for achieving zero inflation or impose a penalty for deviations from it.[18] A third possibility is to appoint a central banker who is known to be "conservative" in the sense of attaching a sufficiently higher marginal disutility (compared with society) to inflation than to output (i.e. a sufficiently lower value of λ in the preceding utility functions). It is also important to protect the central banker from public and governmental pressures to pursue monetary policies (e.g. financing a fiscal deficit) that will cause inflation. This raises the question of the independence of the central bank from the government, discussed earlier in this chapter. Since expectations on central bank policies are important, the central bank's reputation and credibility in maintaining a low inflation policy also become relevant. Walsh (2003) provides additional discussion of the relevant models on these issues. The next section relates intertemporal optimization and commitment to the credibility of monetary policy.

12.5.2 Credibility and the costs of disinflation under the EAPC

Credibility becomes especially important in a situation where the actual and expected inflation rates have become high but the central bank decides to pursue a disinflationary policy (i.e. one which reduces the inflation rate) for the current and future periods and announces a target value of π^{T}. Under full credibility, which requires $\pi = \pi^{\mathrm{T}}$, as the actual inflation rate is cut by contractionary policies, π^{e} adjusts fully and instantly to $\pi^{\mathrm{T}} (= \pi)$, so that $(\pi - \pi^{\mathrm{e}})$ remains at zero and output remains at its full-employment level. With zero credibility, π^{e} would not adjust at all, so that as π is decreased below π^{e} output will fall. Intermediate cases of less than full credibility but more than zero will also be accompanied by some loss of output. Therefore, less than full credibility imposes a loss of output in a period of disinflation.

To relate the above to historical experience, money supply targeting in the late 1970s and 1980s failed to contain inflation to low rates. Among the reasons cited for this failure of the central banks of the USA, UK and Canada, one is that monetary targeting was not pursued vigorously and maintained long enough. Further, the central banks shifted between targets or pursued multiple targets, often overshot their target inflation rate without later reversing the overshoot, and obscured their targets or the reasons for their failure. Added to these was the public's lack of credibility in the central bank's commitment and/or ability to achieve low inflation rates. Another reason for this failure of money supply targeting was the increasing instability of the money demand function, so that the relationship between inflation, output

17 This is rarely done. The European Central Bank, mandated to follow the sole objective of a zero inflation policy, provides a rare example.

18 This would introduce a third term in the objective function, specifying the incentive or penalty. Such a term is usually made quadratic. This discussion is related to proposals for a contract for the central banker.

and money supply had become unstable. Once this became known in the 1980s, money supply targeting was replaced by interest rate targeting.

The 1990s started with the attempts of many countries, including the USA, Canada and the UK, to lower their rates of inflation. However, the public's previous experience was with the higher rates of inflation in the 1980s and the general failure of earlier announcements by their central banks of a lower inflation target rate. The credibility of such targets was low, with the consequence that the disinflation process resulted in recessions and loss of output. Therefore, a pertinent issue for the then debate on the appropriate disinflation policy became whether or not establishing credibility could be beneficial by itself and what losses might flow from the lack of credibility. The following parts of this section explore these issues.

To illustrate the issue of the credibility of monetary policy in the context of the Friedman–Lucas supply constraint, assume that the economy starts with identical and high actual and expected rates of inflation, and that the central bank decides to lower the actual rate of inflation. To achieve this, it follows a restrictive monetary policy, and announces its new policy and the lower inflation rate consistent with that policy. One can imagine several scenarios. In the best of these, the public believes the central bank and immediately lowers its expected inflation rate to the now lower actual inflation rate, so that the economy maintains full employment. But, in an alternative scenario, if the central bank has gone along this route before but not, in the public's experience, delivered the requisite contractionary monetary policy and the announced lower rate of inflation, the public would not find the announcement credible and would not lower its expected inflation rate correspondingly. Consequently, if the central bank did in fact deliver the lower announced actual rate of inflation, the expected rate would exceed the actual one, so that output would fall below the full-employment level. The lack of credibility was the cause of this fall. By extension, greater rather than lesser credibility of the central bank would mean a lower shortfall in output. This argument is sometimes used to assert that inflation, possibly even hyperinflation,can be eliminated from an economy without severe reductions in output and employment – provided that the credibility of the central bank is first established by the appointment of appropriate central bank governors with an established and tough reputation for espousing and establishing price stability.

Consequently, the poor credibility of the central bank introduces a lag in the full impact of monetary policy on the economy, and requires the implementation of stronger policies than would be otherwise necessary in the interim. The extent of the dilution of its impact and the extension of the length of the lag depend on the "extent of credibility" of the central bank's policies.

Credibility and gradualist versus cold turkey policies

Starting from a period of high inflation and low credibility, the restoration of credibility depends on the vigor with which the anti-inflationary policies are pursued. The usual analyses on this issue rest on the propositions that the expected inflation rate adjusts to the actual inflation rate with a lag, so that $\pi > \pi^e$ during the disinflation period, and the adjustment lag of π^e is shorter the more severe the recession in output accompanying the disinflation.

The two extreme alternative ways of curbing high inflation are "gradualism" and "cold turkey" policies. Gradualism is a policy of relatively slow reduction in money growth and inflation rates in an attempt to keep the disruption to aggregate demand and output to a minimum during the adjustment to the low desired inflation rate. The economy is thereby not forced into a more severe recession or a rate of unemployment much higher than the natural one, but the inflation rate falls slowly and takes a longer period to reach the target value.

In the alternative scenario of the cold turkey policy, the money supply growth rate and aggregate demand are cut drastically, causing the economy to go into a more severe recession and raising unemployment substantially. The inflation rate falls rapidly. Given such a strong jolt to output and inflation, the public also adjusts rapidly its expectation of inflation to the central bank's target.

Of the two types of policies, the cold turkey solution is clearly the more forceful and drastic policy and its impact is often more rapid in lowering the expected rate of inflation, as well as in creating credibility that the lower rates will be maintained. This faster reduction in the expected inflation rate means a faster reduction in the actual inflation rate and a reduced lag in the impact of the anti-inflationary policy, which in turn translate into an earlier return to full employment. As against these benefits, the cold turkey policy usually starts by creating a deeper recession and greater rise in unemployment than a gradualist policy. Countries tend to follow different combinations of these two policies at different times.

On the costs of disinflationary policies, Ball (1993) reports for his sample a 0.8 percent reduction, on average, of output below its trend level for every percentage point reduction in inflation. Further, he finds that the output costs of anti-inflationary policies tend to be larger if inflation is reduced gradually over a period of time rather than more rapidly. The cost is likely to be lower if the policy is credible than if it is not credible, and lower under full credibility than under merely rational expectations (Brayton and Tinsley, 1996).

12.5.3 Gains from credibility with a target output rate greater than y^f

The credibility of the central bank's announced inflation rate can yield real gains to the economy, as we illustrate in the following model. This model certainty and implicitly assumes that the central bank has the same utility function as society and that it will achieve the inflation rate that it chooses.

Assume that the central bank wants to choose the optimal values of y and π for its utility function U^M:

$$U^M = -\gamma\pi^2 - [y - ky^f]^2 \qquad \gamma > 0, k > 1 \tag{48}$$

where the symbols are as defined earlier. (48) specifies a loss in utility from inflation, as well as from deviations from the central bank's pre-specified target output ky^f, set in this equation as a *proportion* of the full-employment output y^f. The marginal loss in utility from a unit increase in output deviation has been normalized at unity.

Again, let the economy's aggregate supply relationship be of the Friedman–Lucas type, which is:

$$y = y^f + \alpha(\pi - \pi^e) \qquad \alpha > 0 \tag{49}$$

Substituting (49) in (48) gives:

$$U^M = -\gamma\pi^2 - [(y^f + \alpha(\pi - \pi^e) - ky^f]^2 \qquad \gamma > 0 \tag{50}$$

To find the gains from credibility, we compare the following two extreme cases.

A. There is no credibility and the central bank has to assume the public's expected inflation rate π^e as exogenously given.

For this case, maximize (50) with respect to π for given π^e. This yields the central bank's choice of π as:

$$\pi = \frac{\alpha}{\alpha^2+\gamma}(ky^f - y^f) + \frac{\alpha^2}{\alpha^2+\gamma}\pi^e \tag{51}$$

so that the central bank sets its optimal inflation policy to be a function of the expected inflation rate, the full employment output and the target output. Hence, its target inflation rate will vary as these variables change over time. The inflation rate is positively related to both π^e and $(k-1)y^f$.

Substituting (51) in (49) yields:

$$y = y^f - \frac{\alpha^2}{\alpha^2+\gamma}(1-k)y^f - \frac{\alpha\gamma}{\alpha^2+\gamma}\pi^e \tag{52}$$

Hence, actual output will depend on the target factor k and the expected inflation rate. If $k \neq 1$, y will not equal y^f, even if $\pi^e = 0$. If $k = 1$ and $\pi^e = 0$, $y = y^f$. Further, output is lower for higher values of π^e. Output can differ from its full-employment level but note that, even if the former is higher (in the given objective function, higher than ky^f), there occurs a loss in utility because of the nature of the assumed quadratic objective function.

B. To see the relative gains from full credibility, now assume that the central bank announces the per-period rate of inflation that it intends to achieve *and* the public fully believes it. In this case, $\pi^e = \pi$.

For $\pi^e = \pi$, (49) implies that:

$$y = y^f \tag{53}$$

Therefore, there will not be a loss in output from a disinflationary policy or a gain from an inflationary one, so that, even in the short run, there will not be deviations from full employment and their implied loss in utility. Further, the economy enforces $k = 1$.

Note also that, under full credibility, $y = y^f$ whether the central bank's chosen inflation rate is zero or some positive number. Hence, the output gains from full credibility do not require a zero inflation rate, as long as it is known and fully credible. However, with $\pi^e = \pi$, $y = y^f$, so that $k = 1$. The objective function (48) becomes:

$$U^M = -\gamma\pi^2 - [(y^f - ky^f]^2 \qquad \gamma > 0 \tag{54}$$

Maximizing (54) with respect to π yields the condition:

$$\partial U/\partial\pi = -2\gamma\pi = 0 \tag{55}$$

which gives the optimal rate of inflation π^* as zero. Therefore, for the above model, while the gains from credibility occur at any inflation rate, the optimal inflation policy under full credibility is that of zero inflation. A positive rate of inflation, even though it may be fully credible and maintain full-employment output, will yield a lower utility than zero inflation.

Comparing the full credibility results (with $y = y^f$) with those of no credibility, the central bank can produce higher output by having $k > 1$ and $\pi > \pi^e$. But such deviations from y^f

imply a loss in utility. Further, inflation is positive if there is no credibility and zero if there is, with the former giving lower utility. Hence, full credibility yields higher utility, lower inflation and smaller deviations in output from the full-employment level.

12.5.4 Analyses of credibility and commitment under supply shocks and rational expectations

For further exposition of the modeling of credibility, we modify and extend the preceding model by again bringing in supply shocks and rational expectations. The model used will be:

$$U^M = -\gamma\pi^2 - [y - (y^f + k)]^2 \qquad \gamma, k > 0 \tag{56}$$[19]

$$y = y^{\mathrm{f}} + \alpha(\pi - \pi^{\mathrm{e}}) + \mu \qquad \alpha > 0 \tag{57}$$

where μ is a random disturbance to aggregate supply. Note that the objective function (56) differs from that in (48): the output target level now enters the utility function through an *additive* rather than a proportional factor.[20]

We investigate three scenarios for this model.

I. The central bank does not guarantee that the actual inflation rate will be zero but does pre-commit itself to the average inflation rate π being zero. The public believes it and bases its expectations on this average rate, so that $\pi^{\mathrm{e}} = \overline{\pi} = 0$.

Incorporating this restriction into (57) yields:

$$y = y^{\mathrm{f}} + \alpha\pi + \mu \tag{58}$$

The central bank chooses π to maximize (56) subject to (58), which gives:

$$\pi = -\beta\mu + \beta k \tag{59}$$

$$y = y^{\mathrm{f}} + \alpha\beta k + (1 - \alpha\beta)\mu$$

where $\beta = \alpha/(\alpha^2 + \gamma)$. Hence:

$$y = y^{\mathrm{f}} + \frac{\alpha^2}{\alpha^2 + \gamma} k + \frac{\gamma}{\alpha^2 + \gamma}\mu \tag{60}$$

However, from (59):

$$E\pi = \beta k > 0 \quad \text{for } k > 0$$

19 To allow for a divergence of the central bank's utility function from society's, one possibility would be to define the central bank's objective function as $U^{\mathrm{M}} = S - \gamma\pi^2 - [y - (y^{\mathrm{f}} + k^{\mathrm{M}})]^2$, $\gamma, k^{\mathrm{M}} > 0$, where S is, say, the remuneration of the governor of the central bank, and society's utility function as $U^{\mathrm{S}} = -\gamma\pi^2 - [y - (y^{\mathrm{f}} + k^{\mathrm{S}})]^2$, $\gamma, k^{\mathrm{S}} > 0$. Such divergence can be used to study performance contracts for central banks. If $k^{\mathrm{M}} \neq k^{\mathrm{S}}$, the central bank has a different output target from society and, therefore, a different "agenda" (Waller, 1995).

20 The only justification for this change in the objective function from (48) is our desire to present some variation in modeling. As an exercise, the student should derive the results for this section from (48).

The central bank has given a commitment to $\overline{\pi} = 0$. To meet this commitment when $\beta \neq 0$, it will have to make $k = 0$ or it would be attempting to fool the public and will not be able to maintain its credibility. We designate the credible version (with $k = 0$) of this case as (I′). That is:

I′. Case I above with $k = 0$, so that the target output level is y^{f}.

Note that the requirement of credibility imposes realism on the objectives of the central bank. For this credible policy, $k = 0$, so that:

$$y = y^{\mathrm{f}} + \frac{\gamma}{\alpha^2 + \gamma}\mu \tag{61}$$

$$\pi = -\frac{\alpha}{\alpha^2 + \gamma}\mu \tag{62}$$

Note that for $k = 0$, only the rational expectation of inflation $E\pi$ is zero, so that the public must be taken to accept that actual inflation will vary over time but maintain its expected value as zero. Further, the expectation of output is stabilized at y^f, but its actual level will fluctuate with the actual inflation rate.

II. The central bank guarantees that the actual (rather than the average) inflation rate will be zero. The public believes it and bases its expectations on it, so that $\pi^{\mathrm{e}} = \pi = 0$.

Incorporating this expectations formation into (57) yields:

$$y = y^{\mathrm{f}} + \mu \tag{63}$$

This implies that the central bank cannot achieve $k > 0$. It must, therefore, give up either its target of output greater than the full-employment level or its commitment to a zero inflation rate.

Further, substituting (63) into the objective function (56) gives:

$$U^{\mathrm{M}} = -\gamma\pi^2 - (\mu + k)^2 \qquad \gamma, k > 0$$

Maximizing this function with respect to π gives:

$$\partial U^{\mathrm{M}}/\partial\pi = -2\gamma\pi = 0$$

so that:

$$\pi = 0 \tag{64}$$

Hence, the optimal inflation rate is zero. However, there may be problems in achieving it since the model includes random supply shocks that are unpredictable and the central bank may not be able to implement the policies that will counter their impact on the inflation rate. If the central bank can, in fact, achieve this rate, its optimal inflation policy will be consistent with its commitment to zero inflation.

III. The central bank takes π^e as given for its decisions while the public forms its expectations rationally.

For given π^e, maximizing (56) with respect to π and subject to (57) yields:

$$\pi = \beta(\alpha\pi^e + k - \mu) \tag{65}$$

where $\beta = \alpha/(\alpha^2 + \gamma)$. Assuming rational expectations, so that $\pi^e = E\pi$,

$$\pi^e = (\alpha/\gamma)k \tag{66}$$

Equations (65) and (66) imply that:

$$\pi = \frac{\alpha}{\gamma}k - \frac{\alpha}{\alpha^2 + \gamma}\mu \tag{67}$$

Since $E\pi = (\alpha/\gamma)k$, $E\pi > 0$ for $k > 0$. Therefore, the equilibrium level of output under rational expectations will be:

$$y = y^f + \frac{\gamma}{\alpha^2 + \gamma}\mu \tag{68}$$

where $Ey = y^f$ but y differs from y^f.

We can now compare inflation and output for the three cases. For this comparison, we will use for (I) not the general case ($k \geq 0$) but (I′), which has a credible commitment to zero average inflation ($k = 0$). Ey is identical among them. The fluctuations in y are also identical between (I′) and (III). A similar set of implications also holds for the inflation rate. This is an interesting result into the nature of rational expectations. In (I′), there is a pre-commitment to a credible zero inflation on average. In (III), there is no pre-commitment to any particular rate of inflation but the public, because of its rational expectations, knows and bases its expected rate on the expected value of the inflation rate adopted by the policy maker. Our conclusions from the above analyses are:

- The identity of results between (I′) and (III) shows that rational expectations act as if there were a credible pre-commitment to an average inflation rate, so that it replaces the need for a pre-commitment. As a corollary, to achieve the benefits of zero inflation, the policy maker does not have to give a pre-commitment as long as it sticks with a systematic zero inflation policy.
- A credible policy is more than just informing the public of the intended policy. It has to be realistic or it will soon cease to be credible. This requirement imposes limitations on the central bank's objectives, which in (I′) ensured that the target output was the full-employment output itself, rather than a higher amount.[21]

21 Boschen and Weise (2003) provide evidence that inflations in the 1960s, 1970s and 1980s in OECD countries were often due to policy makers' short-term pursuit of high real growth rates – and there was evidence of a link between inflation and national elections.

- Comparing (I′) and (III) with (II), the expected values of both inflation and output are the same in all cases but their fluctuations differ. Since $\gamma/(\alpha^2 + \gamma) < 1$, output fluctuates less and inflation more under (I′) and (III) than under (II). Hence, rational expectations reduce the variability of output compared with a policy of zero inflation per period.

Finally, some *caveats* to the above conclusions. Some of the conclusions are specific to the assumptions of the above models. Note especially two of these assumptions. One, the assumed quadratic objective function attributes negative utility to even positive deviations of output from its full-employment level.[22] Intuitively, one considers higher output to be more desirable, so that such positive deviations should have been assigned positive utility. Two, the economy's constraint has been assumed to have the form of the Friedman–Lucas supply function (i.e. the expectations-augmented Phillips curve). This belongs in the classical paradigm, whereas the supply constraint of the Keynesian paradigm has a different form. We will leave this differentiation to questions at the end of this chapter and to Chapter 15.

Brayton and Tinsley (1996) provide some evidence on the benefits of credibility. They find that the costs in terms of output of disinflation are higher under imperfect credibility than under full credibility. Further, they are lower under full credibility than under rational expectations.

12.6 Does the central bank possess information superiority?

The success of monetary policy, and its resulting credibility, will depend on knowledge of the structure of the economy and of the future value of output, prices/inflation, etc., at the time the policy impacts on the economy. These are practical questions. On the latter, since there are long lags, say of six quarters, in the impact of monetary policy, the central bank must be able to forecast accurately enough the values of the relevant variables that far in the future. Such forecasts are prone to error. A more limited requirement of a useful role for the central bank is that it should possess an informational advantage over the public on forecasting at lag lengths relevant to monetary policy. This can result from superior information gathering and its evaluation of the evidence. Some studies have found evidence that the US Federal Reserve does possess such superiority (Peek *et al.*, 2003).

12.7 Empirical relevance of the preceding analyses

The analyses of this chapter have assumed special forms of the central bank's objective functions and the constraints on choices imposed by the economy. Neither of these may be valid for any economy or all economies. While the accuracy of the objective function is more difficult to judge from convincing empirical evidence, the empirical

22 This utility function is more in tune with the classical paradigm since it makes the desirable target of monetary policy the full-employment level rather than some higher level. A Keynesian might prefer a utility function that allocates positive utility to positive deviations from full employment.

estimation of the constraints is quite common. In particular, the aggregate supply equation used in much of this chapter has been the Friedman–Lucas supply function in the form of the expectations-augmented Phillips curve. This function does not yield implications consistent with many, or most, of the stylized facts in Chapters 1 and 14 relating money/inflation to output/unemployment. Therefore, though it has been extensively used in this chapter and in much of the relevant literature, it is quite likely to be inaccurate as a specification of the short-run supply constraint imposed by the economy on the central bank's choices. This would affect the validity and value of the implications drawn in this chapter.

However, it does seem that certain policy recommendations can still be maintained. Among these are:

- Intertemporal maximization and its conclusions are preferable to myopic maximization.
- The Phillips curve is not a sufficiently valid description of the economy's short-run supply constraint, so that the policies derived from it would have an inflationary bias and be especially suspect.
- Credibility of the central bank is important to the success of its policies, especially when it wants to reduce inflation. Given the suspect nature of knowledge of the supply constraint, what may be important for the credibility of the central bank is likely to be the maintenance of its objectives over time, rather than of precise policies derived from inaccurate or imprecise constraints.
- Similarly, given the lack of accurate knowledge of the constraints imposed by the economy in both the current and future periods, as well on the actual formation of expectations, time-consistent policies may in practice turn out not to be superior to reoptimization policies with an unchanged objective function.

However, it is likely that fine details on the extent of the gains from credibility, especially when there is no difference in output and employment, will not hold up when the objective function and the specification of the economy's constraints are modified to other forms. Nor will fine details on the superiority of time-consistent policies relative to reoptimization ones.

Conclusions

Policy makers need to choose among multiple goals and take account of constraints imposed by the economy on their choices. Since the government and the central bank are likely to possess different preferences among goals and view these constraints differently, there exists a potential for conflict between the desired goal levels and the appropriate policies to be pursued. This potential for conflict tends to be greater when the economy is displaying high rates of both inflation and unemployment and tends to be low when the economy has low rates of both these variables. Such conflicts between central banks and governments were more evident during the stagflations of the 1970s and 1980s than they have been since then, partly because of the realization that there is no long-run benefit from high inflation rates.

Empirical studies have shown that economies with effectively independent central banks tend to have lower rates of inflation. Therefore, economies with low rates of inflation as a societal objective should maintain the independence of their central banks.

The time consistency debate has clearly established the superiority of monetary policies that are based on intertemporal optimization as against myopic policies (based on one-period optimization at a time), with the latter being either arbitrary or derived from one-period optimization subject to a short-run Phillips curve. The latter procedure has an inflationary bias. Currently, it is not advocated by either the modern classical or the new Keynesian approaches, but was advocated in the 1950s and 1960s by many Keynesians.

Time-consistent policy paths are superior to reoptimization policies if there are no shifts in preferences or the economy's responses over time, as might occur if there is a change in the government or the central bank governor/chairman, if there are shifts in the knowledge of the structure of the economy, or if the central bank's horizon shortens as time passes. Under uncertainty, a time-consistent intertemporal policy path may require elaborate policy rules to be specified *ex ante*. In addition, the benefits of time-consistent optimization over reoptimization are, at best, likely to be small under the more realistic scenario of a horizon that is long and rolling. The reoptimization procedure is likely to become preferable if there are previously unanticipated shifts in the economy's constraints, which are especially likely to occur under uncertainty of the probabilities of future outcomes.

Central banks now accept the conclusion that intertemporal optimization yields better policies than myopic optimization subject to a short-run Phillips curve and that the latter has an inflationary bias. Hence, they accept the essential contribution of the time consistency debate that policies should be formulated with an intertemporal objective function and must take account of the economy's anticipated responses and of the anticipated future consequences of policies. Central banks also accept the caution emerging from this debate against arbitrary changes in monetary policies if there are no significant shifts in knowledge or of the fundamentals of the economy. Beyond these, monetary authorities rarely follow time-consistent policies or their implication of the precise specifications of optimal policies for all future periods and their implementation in those periods.

At the practical level, most central banks maintain discretion to change their policies if the economy changes in a significant manner, and have reoptimization powers to decide the policy actually pursued in each period, with the understanding or legislated requirement that their objectives for the economy do not change. This is currently the practice in the USA, Canada and the UK. Further, limits on the current knowledge of the future course of the economy, such as that on the unemployment and growth rates, are recognized in a cautious approach to monetary policy; central banks usually change the interest rate in small (usually 0.25 percent) steps, leaving open the option of reinforcing this step within a few months, not doing so, or reversing it as the actual state of the economy reveals itself. This is essentially a continual reoptimization procedure, with an unchanged utility function, rather than a strict time-consistent one.

However, no matter what policy is pursued, it is important that the central bank should maintain credibility for its policies. Sticking to pre-announced consistent policies enhances this credibility and the effectiveness of the policies pursued. Frequent shifts in policy reduce this credibility, reduce the effectiveness of the policies and also introduce longer lags in the achievement of the central bank's targets. Credibility requires that the central bank must pursue realistic objectives. If its goal is price stability and it has so announced it, and it operates in the context of a classical-type economy, its target output has to be the full-employment output and not a higher one.

Appendix

Myopic optimal monetary policy without commitment in a new Keynesian framework

This appendix assumes the new Keynesian quadratic objective function and the new Keynesian supply (price/quantity adjustment) equation (rather than the Friedman–Lucas one used in the body of this chapter) as the relevant constraint from the economy's structure. For the following analysis, the new Keynesian objective function is taken to be the quadratic intertemporal objective function:

$$-\tfrac{1}{2}E_t\left\{\sum_{j=0}^{\infty}\beta^j\left(\lambda(x^{\#}_{t+j}-k)^2+\pi^{\#2}_{t+j}\right)\right\} \qquad k\geq 0 \tag{69}$$

where $x^{\#}$ is the output gap ($=y-y^{\mathrm{f}}$), so that the target output level is the full-employment level plus k(i.e. $x^{\#}=y-(y^{\mathrm{f}}+k)$) and $\pi^{\#}$ is the inflation gap, defined as the deviation of inflation from its desired target level π^{T}. β is the central bank's time discount factor and λ is the weight placed by the central bank on the output gap relative to that on the inflation gap.

For *myopic* (one-period) optimization, the intertemporal optimization problem of the central bank can be replaced by maximization *in each period* of the function:

$$-1/2[\gamma(x^{\#}{}_t-k)^2+\pi^{\#2}{}_t]+H_t \qquad K\geq 0 \tag{70}$$

subject to the following new Keynesian supply constraint (new Keynesian Phillips curve) for the current (t) period:

$$\pi_t=\alpha x^{\#}{}_t+h_t \tag{71}$$

where:

$$H_t=-\tfrac{1}{2}E_t\left\{\sum_{j=1}^{\infty}\beta^j\left(\gamma(x^{\#}_{t+j}-k)^2+\pi^{\#2}_{t+j}\right)\right\} \qquad k\geq 0 \tag{72}$$

and:

$$h_t=\beta\pi^{\mathrm{e}}{}_{t+1}+z_t \tag{73}$$

Note that H_t and h_t are given from the perspective of myopic optimization. Treating them as such assumes that expected future output and inflation are independent of the current output gap and inflation, so that they are also independent of the policies pursued by the central bank.[23] The central bank maximizes:

$$-1/2\gamma(x^{\#}{}_t-k)^2-1/2\{(\alpha x^{\#}{}_t+h_t)-\pi^{*}{}_t)\}^2+H_t \qquad k\geq 0 \tag{74}$$

23 This is clearly unrealistic, which contributes to the undesirability of myopic policies.

Its first-order condition is:

$$-\gamma(x^{\#}{}_t - k) - \alpha\{\alpha x^{\#}{}_t + h_t - \pi^*{}_t\} = 0 \qquad k \geq 0 \tag{75}$$

Substituting $\pi_t = \alpha x^{\#}{}_t + h_t$ from the constraint (72) for one of the $x^{\#}{}_t$ in (75) yields:

$$\pi^{\#}{}_t = -\frac{\gamma}{\alpha}(x^{\#}{}_t - k) \tag{76}^{24}$$

Hence, the myopic inflation gap rises with k, which is the amount by which the central bank wants to exceed the full-employment output level. At full employment (*i.e.* $x^{\#}{}_t = 0$), the central bank will obtain a zero inflation rate if it maintains $k = 0$. Conversely, the central bank would ensure deflation by a policy of $k < 0$. Thus, there is a negative tradeoff between current inflation above its target level and the current output gap, which the central bank can exercise by choosing its desired value of k.

Equation (76) can also be stated in the form of a supply equation as:

$$x^{\#}{}_t = -\frac{\alpha}{\gamma}\pi^{\#}{}_t + k \tag{77}$$

where $x^{\#}{}_t$ is the aggregate demand that the central bank has to ensure for the economy. If $\pi^{\#}{}_t > 0$ (i.e. inflation is above target), (77) implies that the central bank reduces aggregate demand, a policy popularized as "*leaning against the wind*."

γ is the weight placed by the central bank on the output gap relative to that on the inflation gap. A central bank that assigns a higher cost to inflation will have a lower value of γ, so that, from (77), it would maintain aggregate demand and output at a lower level and achieve a lower inflation rate.

The central bank can use the money supply or the interest rate as its instrument for achieving the aggregate demand $x^{\#}{}_t$ in (77). If it uses the real interest rate, (77) is substituted in the IS equation to derive the appropriate real interest rate (r_t) to be set by the central bank. This step requires the specification of the actual form of the IS equation, of which there are many forms in the literature. However, for all of them, the central bank reduces aggregate demand by raising the real interest rate, so that if $\pi^{\#}{}_t > 0$ 0 (i.e. inflation is above target) the central bank raises the real interest rate. In perfect markets, the nominal interest rate (R_t) can be deduced from the Fisher equation ($R_t = r_t + \pi^{\mathrm{e}}{}_t$). Therefore, to reduce aggregate demand, the central bank raises the nominal interest rate sufficiently to raise the real interest rate.

If the central bank chooses the money supply as its instrument, it must ensure the money supply implied by the IS–LM analysis is required for aggregate demand. For this, the values of $x^{\#}{}_t$ and R_t are substituted in the money demand equation to derive the required money supply. Since there are many forms of the IS and LM equations in the literature, we do not proceed further with such derivations.

24 Note that assuming rational expectations and substituting for $x^{\#}{}_t$ from the budget constraint (29) yields:

$$\pi^{\#}{}_t = \frac{\gamma(\alpha k + z_t + \beta E_t \pi_{t+1})}{\alpha^2 + \gamma}$$

where the terms involving $E\pi$ can be iterated forward and eliminated, leaving the right side with only parameters. However, the assumption of rational expectations in the context of myopic optimization when the future inflation rates are taken to be independent of the current policy seems questionable.

Intertemporal optimization with commitment in a new Keynesian framework

For intertemporal optimization, the expected future values of the variables cannot be taken as independent of their current values and the current policies pursued. The central bank now maximizes its preference function subject to the set of constraints, with a separate but similar constraint for each period and with values of the variables dependent on the policies pursued in the past and the future. That is, the expected future values $x^{\#}{}_{t+1}$ and $\pi^{\#}{}_{t+1}$ cannot be assumed to be independent of the actually chosen values of $x^{\#}{}_{t}$ and $\pi^{\#}{}_{t}$, and vice versa. The application of rational expectations to such values is appropriate in the intertemporal context.

For this problem, the central bank maximizes, over an infinite sequence of $(x^{\#}{}_{t+j}, \pi^{\#}{}_{t+j})$, $j = 0, 1, \ldots, \infty$, the objective function:

$$-\tfrac{1}{2}E_t\left\{\sum_{j=0}^{\infty}\beta^j\left(\gamma(x^{\#}_{t+j}-k)^2+\pi^{\#2}_{t+j}\right)\right\} \qquad \gamma,\beta,k\geq 0 \tag{78}$$

subject to the set of constraints, one for each period, generated by the economy. These constraints, generalized from the one specified above from the new Keynesian approach and using rational expectations, are:

$$\pi_{t+j}=\alpha x^{\#}{}_{t+j}+\beta E\pi_{t+1+j}+z_{t+j} \qquad \alpha,\beta>0 \tag{79}$$

For this problem, the Lagrangian is:

$$\max\ -\frac{1}{2}E_t\left[\sum_{j=0}^{\infty}\beta^j\{(\gamma(x^{\#}{}_{t+j}-k)^2+\pi^{\#2}_{t+j})+\lambda_{t+j}\pi_{t+j}-\alpha x_{t+j}-\beta\pi_{t+1+j}-z_{t+j}\}\right] \tag{80}$$

Under the assumption that output cannot be maintained on average at a level above its full-employment level, k can be set at zero. The solution (derivation not shown) for this problem is:

$$x^{\#}{}_{t}=-\frac{\alpha}{\gamma}\pi^{\#}{}_{t} \tag{81}$$

$$x^{\#}{}_{t+j}-x^{\#}{}_{t+j-1}=-\frac{\alpha}{\gamma}\pi^{\#}{}_{t+j} \qquad \text{for all } j\geq 1$$

These expressions substituted in the IS equation, yield the interest rate policy, which when substituted in the LM equation yields the money supply policy.

Summary of critical conclusions

- If the economy does allow short-run tradeoffs among multiple goals, there is a high potential for periodic conflicts among policy makers in the goals attempted and the policies pursued.

- ❖ If the economy does not allow monetary policy to affect output and unemployment even in the short run – that is, if money is neutral – the adoption of price stability or a low inflation rate as the single or dominant monetary policy goal becomes more clearly the optimal policy goal. It also reduces the potential for conflict between monetary and fiscal policies.
- ❖ Central bank independence of the government has been found to reduce the rate of inflation.
- ❖ Time-consistent policies need not be superior to those derived from intertemporal optimization with an unchanging objective function and a long rolling horizon.
- ❖ Both types of intertemporally optimal monetary policies are superior to myopic policies and buttress the central bank's credibility.
- ❖ The credibility of the central bank is essential to its successful reduction of inflation rates. Credibility is also a factor in reducing the time lags in the adjustment of the expected inflation rate and of the actual inflation rate.
- ❖ The credibility of a policy committed to keeping the price level stable imposes a requirement that the central bank must not try to achieve a target output higher than the full-employment level.

Review and discussion questions

1. Who should determine the economic policy goals for the nation: the government democratically elected by the public or a central bank whose directors (or governors) are not elected and cannot be made directly responsible to the public? What are some of the practices in this respect?
2. It is often argued that independence of the central bank from the government and the legislature would lead to a lower inflation rate. Why, and under what circumstances is this likely to happen?
3. The time consistency of policy requires the specification of a policy plan for the future, say for the next five years, and the determination to stick to it. Under what conditions – and shocks – is this a desirable policy? When is it not desirable? Discuss.
4. Show that there can be an inflationary bias in discretionary myopic policies.
5. What is the relationship between rational expectations and the credibility of monetary policy? What do they imply for the relevance of credibility to the success of an anti-inflation program? What do they imply for the choice between a gradualist versus a cold-turkey approach to fighting inflation? Discuss.
6. "The time consistency literature suggests that it is always a good thing to have a tough central banker who cares about inflation and not at all about unemployment – even if society cares mainly about unemployment." Discuss.
7. "For pre-specified goals, it is easy for the academic economist to prescribe the course of monetary policy for the period ahead. But the devil is in the details: the limitations on the knowledge of the future course of the economy and its future response to policies, as well as in the implementation of the chosen policies. It is such details that make the successful pursuit of policies an art rather than a science." Discuss.
8. "Each time the central bank changes its discount rate, there always seems to be plenty of criticism from many economists, some claiming that the change is not needed or too much while others claim that it is too little. Is this due to the argumentative disposition of economists as human beings, as the public seems to suspect, or to the nature of their discipline, or are there other reasons?" Discuss.

9. Present the IS–LM analysis for the case of negative shifts in the vertical aggregate supply curve when the targets are (i) aggregate demand, (ii) price stability, (iii) output. Compare the output and interest rate fluctuations under these targets.
10. Assume that the central bank has the utility function:

$$U = -u^2 - 0.5\pi^2$$

where U is utility, u is unemployment and π is the inflation rate. Assume that the economy imposes a constraint in the form of the Phillips curve. Let this be:

$$u = 5 - 0.1\pi$$

Derive the optimal values of u and π. Compare these optimal values with those under a credible pre-commitment to zero inflation.
11. Given the utility function in the preceding question, assume that the economy has an expectations-augmented Phillips curve such that:

$$u - u_n = -0.1(\pi - \pi^e)$$

Derive the optimal values of u and π for (a) exogenously given π^e, (b) the rationally expected value of π^e. Compare these optimal values with those under a credible pre-commitment to zero inflation.
12. Suppose that the economy is such that a positive monetary shock reduces unemployment. Assume that the central bank likes a reduction in unemployment but dislikes an increase in inflation. The public forecasts money growth from the government's optimization problem. How are money growth and inflation determined in this context? If you so wish, you can answer this by specifying a model. What would be the implications of a commitment – i.e. a rule – binding the central bank to a future rate of money growth? Should it do so?
13. What advice would you give the central bank of your country on the appropriate monetary policy to follow at the present time? Your answer must specify your assessment of the current state of the economy, and the goals and targets you think the central bank should follow and how they agree with or deviate from those being followed by the central bank. You must present a detailed analysis of the impact of your suggested policy on output, inflation and interest rates.

References

Alesina, A., and Summers, L.H. "Central bank independence and macroeconomic performance: some comparative evidence." *Journal of Money, Credit and Banking*, 25, 1993, pp. 151–62.

Ball, L. "How credible is disinflation? The historical evidence." *Federal Reserve Bank of Philadelphia Business Review*, 1993, pp. 17–28.

Ball, L. "Policy rule for open economies." In J.B. Taylor, ed. *Monetary Policy Rules*. Chicago: University of Chicago Press, 1999.

Barro, R.J., and Gordon, D.B. "Rules, discretion and reputation in a model of monetary policy." *Journal of Monetary Economics*, 12, 1983, pp. 102–21.

Blackburn, K., and Christensen, M. "Monetary policy and policy credibility: theories and evidence." *Journal of Economic Literature*, 27, 1989, pp. 1–45.

Boschen, J.F., and Weise, C.L. "What starts inflation: evidence from the OECD countries." *Journal of Money, Credit and Banking,* 35, 2003, pp. 323–49.

Brayton, F., and Tinsley, P. "A guide to FRB/US: a macroeconomic model of the United States." *Federal Reserve Board Finance and Economic Discussion Series*, 1996–42, 1996.

Clarida, R., Gali, J. and Gertler, M. "The science of monetary policy: a New Keynesian perspective." *Journal of Economic Literature*, 37, 1999, pp. 1661–707.

Cukierman, A., Webb, S.B. and Neyapti, B. "Measuring the independence of central banks and its effects on policy outcomes." *World Bank Research Review*, 6, 1993, pp. 353–98.

Eijffinger, S.C.W., and de Hahn, J. *The Political Economy of Central Bank Independence*. Princeton NJ: Princeton University, Special Papers in International Economics, 19, 1996.

Fischer, S. "Rules versus discretion in monetary policy." In B.M. Friedman, and F.H. Hahn, eds, *Handbook of Monetary Economics*, vol. II. Amsterdam: North-Holland, 1990.

Goodhart, C.A.E. "Central bank independence." *Journal of International and Comparative Economics*, 3, 1994. Also in his *The Central Bank and the Financial System*. London: Macmillan, 1995.

Kydland, F.E., and Prescott, E.C. "Rules rather than discretion: the inconsistency of optimal rules." *Journal of Political Economy*, 85, 1977, pp. 473–93.

Lucas, R.E., Jr. "Nobel lecture: Monetary neutrality." *Journal of Political Economy*, 104, 1996, pp. 661–82.

Mankiw, N.G. "The inexorable and mysterious tradeoff between inflation and unemployment." *Economic Journal*, 111, 2001, pp. C45–C61.

Peek, F., Rosengren, E. and Tootell, G. "Does the Federal Reserve possess an exploitable information advantage?" *Journal of Monetary Economics*, 50, 2003, pp. 817–39.

Rotemberg, J., and Woodford, M. "Interest rate rules in an estimated sticky price model." In J.B. Taylor, ed., *Monetary Policy Rules*. Chicago: University of Chicago Press, 1999.

Waller, C.J. "Performance contracts for central bankers." *Federal Reserve Bank of St Louis Review*, 1995, pp. 3–14.

Walsh, C.E. *Monetary Theory and Policy*, 2nd edn. Cambridge, MA: MIT Press, 2003.

Woodford, M. "How important is money in the conduct of monetary policy?" *NBER Working Paper* no. 13325, 2007.

Part V

Monetary policy and the macroeconomy

13 The determination of aggregate demand

This chapter derives the aggregate demand for commodities in the economy from analyses of the commodity and money markets. Depending on the central bank's choice of its primary monetary instrument, there are two variants of the money market analysis. In one variant, the central bank controls the money supply in a manner that makes it exogenous to the economy. In the other variant, the central bank controls the interest rate in a manner that makes it exogenous to the economy. The first variant leads to the IS–LM analysis of aggregate demand. The second one leads to the IS–IRT analysis of aggregate demand. The selection between these models for a particular economy depends on the empirical question: does the central bank make the money supply or the interest rate exogenous to the economy?

In both models, monetary and fiscal policies are effective in changing aggregate demand. However, the addition of Ricardian equivalence to both models makes aggregate demand invariant to fiscal policy changes.

Key concepts introduced in this chapter

- IS relationship/curve
- LM relationship/curve
- IS–LM analysis
- Interest rate targeting (IRT)
- Taylor rule
- IS–IRT analysis
- Aggregate demand relationship/curve
- Ricardian equivalence
- St Louis equation

This chapter sets out the two variants of the model of aggregate demand that serves as the workhorse of short-run macroeconomic analysis. The first variant is the IS–LM model, which assumes that the central bank exogenously sets the money supply, and the second variant is the IS–IRT model, which assumes that the central bank exogenously sets the interest rate.

The IS–LM and IS–IRT models aggregate the very large number of goods in the economy into four types: commodities, money, bonds and labor, with an additional good, "foreign exchange," for the open economy. Of these goods, money is defined as that good which

serves as a medium of payments and bonds are defined as "non-monetary financial assets," so that they include what are known in everyday parlance as bonds, loans and stocks. The macroeconomic models analyze the markets for these goods and are basically those oriented towards the study of general fiscal and monetary policies and their implications for aggregate output, employment, interest rates and prices in the short run. The short run in this context is defined as the analytical (not chronological) period during which the capital stock, labor force and technology are exogenously given by assumption.

The basic short-run macroeconomic model is known as the IS–LM model if the money supply, rather than the interest rate, is exogenously set by the central bank. The corresponding model when the interest rate, rather than the money supply, is assumed to be exogenously set by the central bank, can be designated as the IS–IRT model. The choice between the IS–LM and IS–IRT models for the determination of aggregate demand for a given economy is not a matter of dogma or school but depends on the answer to the empirical question: which is the exogenous monetary policy variable and which is the endogenous one? Therefore, before proceeding to the selection of one of these models for a given economy, the appropriate primary operating target of monetary policy used by the central bank needs to be determined. Hence, the model relevant to one economy can differ from that of another economy if the appropriate operating target is different.

Note that the complete short-run macroeconomic model needs both the aggregate demand and aggregate supply function for commodities in order to determine the price level and aggregate output and employment. The determination of aggregate supply is left to Chapter 14 for the classical group of models and to Chapter 15 for the Keynesian group.

13.1 Boundaries of the short-run macroeconomic models

The basic orientation of the short-run macroeconomic models is towards a study of the impact of the general fiscal and monetary policies upon aggregate output and the general price level. There is no attempt to examine their impact upon the relative prices of commodities, so that these models operate as if in a single commodity world. Further, labor is treated as if it were a homogeneous input. It is also assumed that the existing capital stock is given and that while the purchase of commodities for additions to the capital stock (i.e. investment) is made, the productive capacity of the stock does not change because of the gestation lags between purchases of equipment and its use for production. We are, therefore, left with only one variable input, labor, in the production process. The fixity of capital is one of the distinguishing characteristics of short-run versus growth models.

The open economy has five basic goods: commodities, money, bonds, labor and foreign exchange. A study of their markets and "prices" requires specification of the demand and supply for each good, and their equilibrium. Their analysis throughout this chapter is one of comparative statics, focusing on equilibrium states but with some discussion of out-of-equilibrium adjustments.

13.1.1 Definitions of the short-run and long-run in macroeconomics

The short run is that analytical period in which some variables in the model are held fixed. Therefore, depending upon which variable is being held fixed, there can be many different designations of the short run.

In specifying the macroeconomic model to be short run, rather than a long-run model, one of the variables being held constant is the physical capital stock. Investment in physical

capital occurs in the model but the implicit assumption is that there is a sufficiently long gestation lag for such investment for it not to change the productive capacity of the economy. However, this investment does change aggregate demand in the economy during the short run. In contrast to the fixity of the physical capital stock in the short-run models, growth models assume that investment changes the physical capital stock.

Another type of fixity that occurs within the short-run macroeconomic models is with respect to expectations under uncertainty. The short run allows the expected price level to differ from the actual price level. This requires a theory on the formulation of the expected price level. Under rational expectations, this error would be random. The definition of the long run includes the restriction that there are no errors in expectations. Therefore, "long run" analysis assumes that expected prices are identical to the actual prices, as well as having a variable amount of physical capital.

13.2 The foreign exchange sector of the open economy and the determination of the exchange rate under floating exchange rates

The balance of payments equals the inflows of foreign exchange from commodity exports, capital imports/inflows and the net inflows of interest and dividend payments and unilateral transfers less the outflows of funds for commodity imports and capital exports. This difference results in corresponding changes in the foreign exchange held in the domestic country. Assuming a flexible exchange rate, the usual assumption for macroeconomic analysis is that the nominal exchange rate adjusts to ensure continuous equilibrium in the balance of payments, so that $B = \Delta FR = 0$.

$$B = \Delta FR = (X_c - Z_c) + (Z_k - X_k) + NR + NT = 0 \tag{1}$$

where:

FR = foreign exchange reserves
B = balance of payments
X_c = value of exports of commodities (goods and services)
X_k = value of (financial) capital exports
Z_c = value of imports of commodities (goods and services)
Z_k = value of (financial) capital imports
NR = net interest and dividend inflows
NT = net unilateral transfers (gifts and donations) to the domestic economy from abroad

Exports and imports depend on the real exchange rate, specified as:

$$\rho^r = \rho P / P^F$$

where:

ρ = nominal exchange rate, defined as units of the foreign currency per unit of domestic currency (e.g. as £ per $)
ρ^r = real exchange rate ($= \rho P/P^F$ = foreign commodities per unit of domestic commodity)
P = domestic price level
P^F = foreign price level

We simplify by assuming that both exports and imports have price elasticities greater than unity. Therefore, our nominal exports X_c decrease as the real exchange rate $\rho^r (= \rho P/P^F)$ rises (which will occur if the domestic price level P rises, ρ rises[1] or/and the foreign price level P^F falls) since our commodities would become relatively more expensive compared with foreign goods. Our exports also increase with an increase in foreign income y^F. Similarly, our nominal imports Z_c increase if ρ^r rises (since a unit of our commodities will buy more of foreign ones so that foreign commodities become relatively less expensive) or if domestic income y rises. Capital flows depend upon a range of factors, of which the rates of return on domestic and foreign assets are likely to be the most important ones. It will be assumed that capital exports X_k decrease, and capital imports Z_K increase, as the domestic nominal interest rate R rises or the foreign interest rate R^F falls.[2] Putting these ideas formally into (1), the equilibrium condition for the foreign exchange market becomes:

$$\underset{- \ +}{[X_c(\rho^r, y^F)} - \underset{+ \ +}{Z_c(\rho^r, y)}] + [\underset{+ \ -}{Z_k(R, R^F, \rho''^e)} - \underset{- \ +}{X_k(R, R^F, \rho''^e)}] + NR + NT = 0 \tag{2}$$

where:

y = domestic national income
y^F = foreign national income
R = domestic nominal interest rate
R^F = foreign nominal interest rate
ρ''^e = expected appreciation of the exchange rate

The signs under the variables indicate the signs of the respective derivatives or elasticities. *NR* (net interest inflows) and *NT* (net inflows of transfers) tend to be largely exogenous. To the extent that they are not exogenous, their dependence on incomes and interest rates can be captured in either commodity or capital flows.

We have assumed in the introduction to this chapter that the domestic economy is *small* relative to the rest of the world: that is, the variables P^F, y^F and R^F would not be affected by changes in our exports and imports of commodities and capital. They are, therefore, exogenous variables whose values are given for our economy. Omitting the variables y^F and R^F from equation (2) since they are exogenous and omitting ρ''^e for simplification, we are left with the equation:

$$\underset{-}{[X_c(\rho^r} - \underset{+ \ +}{Z_c(\rho^r, y)}] + [\underset{+}{Z_k(R)} - \underset{-}{X_k(R)}] + NR + NT = 0 \tag{3}$$

Substituting $\rho P/P^F$ for ρ^r in (3) we get:

$$[X_c(\rho P/P^F) - Z_c(\rho P/P^F, y)] + [Z_k(R) - X_k(R)] + NR + NT = 0 \tag{4}$$

1 Under our definition of the exchange rate, ρ falling (i.e. less £ per $) means that the domestic currency – the dollar – has depreciated, which makes our goods cheaper for foreigners and increases our exports, while making foreign goods more expensive for us and decreasing our imports.

2 It is also assumed that the other endogenous variables such as incomes, prices and exchange rates are not likely to significantly affect capital flows from the viewpoint of a purely comparative static analysis.

This equation – designated the BP (equilibrium) equation – becomes an element of the open economy macroeconomic model. Solving it for the equilibrium value ρ^* of the nominal exchange rate ρ yields:

$$\rho^* = f(P/P^{\mathrm{F}}, y, R;\ \text{other exogenous variables and parameters in the BP equation}) \quad (5)$$

For a country with a flexible exchange rate, the usual assumption for convenience in macroeconomic analysis is that the exchange rate will always be the equilibrium one, thereby ensuring *continuous* balance of payments equilibrium for the given values of R and y. Therefore, under flexible exchange rates, ρ^* from (5) is substituted for ρ in the open economy IS equation to derive the final form of the IS equation and curve.

For the following analysis of the commodity market, we do need to note the relationship between ρ, y and R. Assuming a high level of capital mobility in response to interest rate differentials, an increase in R, with R^{F} held constant, increases net capital flows to the domestic economy, which causes ρ to appreciate, which, in turn, decreases net exports. The decline in net exports decreases expenditures on domestic commodities and reduces y. Therefore, for the open economy, an increase in R, with the expected inflation rate held constant, increases the real interest rate r and decreases y, so that $\partial y/\partial r < 0$. The sign of this effect is the same as for a closed economy, in which an increase in r reduces investment and therefore y. Hence, the open economy maintains $\partial y/\partial r < 0$, but with a larger absolute value than for the closed economy.

Note that an increase in y induces an increase in imports, which decreases net exports and leads to a decline in ρ. That is, $\partial\rho/\partial y < 0$. As R decreases, net capital inflows fall, which causes ρ to decline, so that $\partial\rho/\partial R > 0$.

The preceding analysis is, of course, not relevant if the domestic country maintains a fixed exchange rate. In this case, ρ will be a constant. However, the determination of aggregate demand in this context will require the simultaneous solution of the three equations for equilibrium in the commodity market, in the money market and in the foreign exchange market. This chapter does not present the relevant analysis for the fixed exchange rate case, but does so only for the flexible exchange rate case.

13.3 The commodity sector

Equilibrium in the open-economy commodity market

The real expenditures e on the commodities produced in the open economy, which constitute the aggregate sales revenue of all the firms in the economy, are given by:

$$\begin{aligned} e &= (c - z_c/\rho^{\mathrm{r}}) + i + g + x_{\mathrm{c}} \\ &= c + i + g + (x_c - z_c/\rho^{\mathrm{r}}) \end{aligned} \quad (6)$$

where all the variables are in real terms and:

e = (real) expenditures on domestic commodities
c = total consumption expenditures
$(c - z_c/\rho^{\mathrm{r}})$ = consumption expenditures on domestic commodities

i = (intended) investment[3]
g = government expenditures on commodities[4]
z_c/ρ^r = commodity imports (in domestic real terms).

Our exports are already in domestic commodities and are therefore in domestic real terms. $z_c/\rho^r (= (P^F/\rho P)z_c)$ are expenditures in real terms on imported commodities. The reason for ρ^r in the denominator is as follows. z_c is the *quantity* of imported goods bought at foreign prices P^F (where F stands for foreign), so that £$P^F z_c$ (where £ stands for the foreign money) is our expenditure in the foreign currency on imported goods. This amount has to be converted into the domestic currency by dividing by the nominal exchange rate ρ, so that $\$(P^F/\rho)z_c$ is the domestic or dollar price of the imported goods. This nominal amount in dollars has to be deflated into real terms at the domestic price level P to find the expenditures in domestic real terms on the imported commodities. Therefore, the real value of imported commodities in the above equation is not merely z_c but $(P^F/\rho P)z_c$, which equals z_c/ρ^r.[5]

In the aggregate over all domestic firms, firms' output equals their total payments to the factors of production, which together constitute national income y. National income can be spent by domestic residents on consumption c (which includes the consumption of both domestic and imported goods), on saving s and on payments of net taxes t. Hence,

$$y = c + s + t \tag{7}$$

where:

s = (real) private saving
t = (real) net taxes paid (net of transfers)

The equilibrium condition for the open economy is:

$$e = y \tag{8}$$

In this equilibrium, firms' revenues e and costs of production y are equal, so that they would maintain production at an unchanged level. But if $e > y$, firms' revenues would exceed their costs and they would be tempted to expand production. If $e < y$, firms' revenues would be less than their costs and they would attempt to reduce losses by contracting production.

From (6) to (8), the equilibrium condition for the commodity market can also be stated as:

$$i + g + x_c = s + t + z_c/\rho^r \tag{9}$$

$$s = i + (g - t) + (x_c - z_c/\rho^r)$$

Saving is the residual of current output left over after consumption and payment in taxes. This residual may be directly committed for future production in the form of investment,

3 This mode of defining investment excludes "unintended investment," which mostly consists of commodities produced in the current period and intended for sale but remaining unsold by the end of the period, thereby becoming part of the physical capital stock. The total or accounting measure of investment is the sum of investment as defined above and unintended investment.

4 Note that government expenditures exclude subsidies and other transfer payments to the public (including firms).

5 Some authors do not divide z_c by ρ^r under the assumption that ρ^r always equals unity or that this is a plausible enough assumption for the short run. This is done under the implicit assumption that purchasing power parity (PPP) holds, which makes ρ^r equal to unity. However, PPP does not normally hold, even over periods as long as ten or twenty years, let alone periods as short as a month. Therefore, we do not assume that $\rho^r = 1$.

lent to others in exchange for bonds, exchanged for money balances or sent abroad as net exports. Therefore, even in equilibrium, the overall demand for domestic investment goods is unlikely ever to equal the domestic supply of saving in an open economy in which other goods, such as bonds, money and foreign goods, exist.

Equilibrium does not always exist in the commodity sector, so that the commodity market can have excess demand or supply for some time. Such an excess demand or supply may be eliminated relatively soon or with a considerable time lag and depends upon the speed of adjustment towards equilibrium.

13.3.1 Behavioral functions of the commodity market

Further analysis of the commodity market requires specification of the functions for each of the variables in the preceding equations. We can proceed further with general forms of these functions or, for simplification, assume linear forms for them. The latter are usually given the forms:

$$c = c(y_d) = c_0 + c_y(y - t) \qquad 0 \leq c_y \leq 1 \tag{10}$$

$$i = i(r) = i_0 - i_r r \qquad i_{\mathrm{r}} \geq 0 \tag{11}$$

$$t = t_0 + t_y y \qquad 0 \leq t_y \leq 1 \tag{12}$$

$$g = g_0 \tag{13}$$

$$x_c = x_c(\rho^{\mathrm{r}}) = x_{c0} - x_{c\rho}\rho^{\mathrm{r}} \qquad x_{c0}, x_{c\rho} > 0 \tag{14}$$

$$z_c = x_c(y_d, \rho^{\mathrm{r}}) = z_{c0} + z_{cy}y_d + z_{c\rho}\rho^{\mathrm{r}} \qquad z_{cy}, z_{c\rho} > 0 \tag{15}$$

where r is the real rate of interest and $y_d(= y - t)$ is disposable income. Exogenous foreign variables have been omitted from the equations. Symbols with subscripts are parameters. We also have the commodity market equilibrium condition:

$$e = y \tag{16}$$

Intertemporal utility maximization subject to the budget constraint implies that consumption depends negatively on the real rate of interest, so that (10) should have included another term such as $-c_r r$ on the right-hand side. For the short run, the empirical dependence of consumption on the rate of interest, for the usual range of interest rates in the developed economies, is doubtful – and definitely much more doubtful than the dependence of investment on the rate of interest. Further, the impact of such a term within the macroeconomic model is identical with that of the dependence of investment on the rate of interest. Consequently, we have not included the rate of interest in the consumption function; by implication, the rate of interest will also not be in the saving function.[6]

6 If we believe that saving depends positively on the rate of interest, this can be easily incorporated into the model by redefining $(-i_{\mathrm{r}})$ to measure the decrease in investment plus the decrease in consumption (or less the increase in saving) induced by the marginal increase in the rate of interest.

Note that the preceding equations assume that consumers and firms are free of price illusion in their consumption and investment decision.[7] This assumption is the standard one in macroeconomic models. Further, the preceding consumption and investment functions are simplified versions of more complicated behavior. In particular, they ignore the dependence of consumption on consumer confidence and of investment on business confidence, which depend on the expectations of future income, availability of jobs and expected demand.

The IS relationship

The preceding equations for the open-economy commodity market imply that:

$$y = \left(\frac{1}{1 - c_y + c_y t_y + \frac{1}{\rho^r} z_{cy}(1 - t_y)} \right) \cdot$$
$$[\{c_0 - c_y t_0 + i_0 - i_r r + g + x_{c0} - x_{c\rho}\rho^r\} + (1/\rho^r) \cdot \{-z_{c0} + z_{cy} t_0 - z_{c\rho}\rho^r\}] \qquad (17)^8$$

This equation is the *IS equation for the open economy*.[9] It makes y a function of r and ρ^r, where $\rho^r = \rho P/P^F$. Replacing ρ^r by $\rho P/P^F$ makes y a function of r, ρ, P and P^F, in addition to the policy and exogenous variables and parameters. As explained in the preceding section on the balance of payments, under flexible exchange rates the equilibrium value ρ^* of the nominal exchange rate is determined by the balance of payments equilibrium. Replacing ρ by the determinants of ρ^* gives the final form of the IS equation. In this equation, y will be a function of P, r, the fiscal policy variables and the exogenous foreign variables P^F, r^F and y^F, but ρ will no longer appear as a separate determinant of y. Since this form is considerably more cumbersome[10] than (17), the standard form of the IS equation, our further analysis of the commodity sector will be based on (17) with the understanding that ρ^r is to be replaced by $\rho P/P^F$ and ρ is replaced by the determinants of its equilibrium value ρ^* from the balance of payments equilibrium condition (4).

However, we do need to derive the sign of $\partial y/\partial r$ for the open economy. Under flexible exchange rates, the earlier balance of payments equilibrium analysis implied that ρ will be a function, among other variables such as P and P^F, of the domestic and foreign nominal interest rates R and R^F. Focusing on the relationship between ρ and R, it was shown that an increase in R, with R^F held constant, decreases y. An increase in R, with the expected inflation rate held constant, also increases r. Therefore, for the open economy, $\partial y/\partial r$ is negative for two reasons: investment is negatively related to r and the increase in r causes an appreciation of ρ and decreases net exports. Hence, the open economy maintains $\partial y/\partial r < 0$, but with a larger absolute value than for the corresponding closed economy.

7 This is not, however, supported by the existing versions of those consumption and investment theories since these theories permit substitution between present purchases of commodities for consumption or investment and those in the future as prices change over time.

8 This is the IS equation without Ricardian equivalence.

9 It was given this name since it was originally derived for a closed economy model without a government sector. Hence, it represented the equilibrium condition $i = s$.

10 In addition, the components of the balance of payments are usually not given commonly accepted simple linear representations, as is the case of the components of aggregate demand, so that it is difficult to find a linear specification of the open economy IS equation that embodies the determinants of the equilibrium exchange rate.

In the IS equation (17), [.] in the numerator has two terms involving the real exchange rate ρ^{r}, with both preceded by a negative sign, so that an appreciation of the exchange rate decreases national income/expenditures y. The first term is $x_{c\rho}\rho^{r}$. Its negative sign occurs because as ρ^{r} increases, exports fall, which decreases the expenditures on domestic commodities.[11] The second term is $z_{c\rho}\rho^{r}$. Its negative sign occurs because as ρ^{r} increases, imports rise, so that the leakages from domestic income (similar to those due to taxes paid) rise, thereby decreasing the share of disposable income that can be spent on domestic commodities.[12]

To emphasize, the preceding remarks imply that ρ is not constant along the IS curve. It falls as we move down the IS curve to a lower level of r and a higher level of y. This occurs for two reasons. One of these is that as y increases, imports rise and net exports fall, so that ρ declines. The second reason is that as r decreases, capital inflows fall, which also causes ρ to decline.

Replacing (.) by α in the above IS equation, the equation can be written as:

$$y = \alpha[\{c_0 - c_y t_0 + i_0 - i_r r + g_0 + x_{c0} - x_{c\rho}\rho^{r}\} + (1/\rho^{r}) \cdot \{-z_{c0} + z_{cy} t_0 - z_{c\rho}\rho^{r}\}] \quad (18)$$

where the symbol α has the meaning:

$$\alpha = \left(\frac{1}{1 - c_y + c_y t_y + \frac{1}{\rho^{r}} z_{cy}(1 - t_y)} \right) > 0 \quad (19)$$

Equation (18) implies that the equilibrium level of real income is a function of the rate of interest and the price level, so that the general form of the open economy IS equation is $y = y(r, P)$. This general form would remain valid even if consumption depended upon income *and* the rate of interest, and if investment depended upon the interest rate *and* income.[13] However, the parameters of (18) – and the multipliers below – would alter in such a case. This general form changes to $y = y(r)$ if purchasing power parity, with ρ^{r} equal to 1 or a constant, is maintained as an assumption. However, the assumption of PPP is not warranted for short-run models, since there are doubts about its empirical validity for very long periods, let alone short ones. This general form would also become $y = y(r)$ if the economy were a closed one, so that the export and import terms would drop out of the IS equation. However, most modern economies are increasingly open ones, with exports (or imports) being a large and growing proportion of their GDP.

Equation (19) is plotted in Figure 13.1 as the IS curve in the (r, y) space. The intuitive explanation for the negative slope of the IS curve is as follows. An increase in income y, along the horizontal axis, increases both consumption and saving, as well as increasing tax revenues. Equilibrium requires that investment must also increase by the combined increases in saving and tax revenues. But this can only occur if the rate of interest falls. Hence, an increase in income must be accompanied by a decline in the interest rate if equilibrium is to be maintained in the commodity market. Note that the open-economy IS curve depends on the real exchange rate and through it on the domestic and foreign price levels. Also, as

11 Therefore, the IS curve will shift to the left and aggregate demand will fall.

12 Therefore, the IS curve will shift to the left.

13 The general and linearized forms of these functions would then be: $c = c(y, r) = c_0 + c_y y - c_r r$ and $i = i(r, y) = i_0 - i_r r + i_y y$, with the assumption that $c_y(1 - t_y) + i_y < 1$.

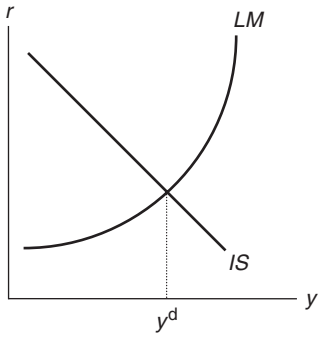


Figure 13.1

shown earlier, the nominal exchange rate ρ declines as the economy moves down the IS curve. But an increase in the domestic price level makes domestic commodities relatively more expensive than foreign ones and decreases net exports, so that expenditures on domestic commodities fall and the IS curve shifts to the left for any given interest rate.

The commodity market (partial) multipliers

From the preceding IS equation, some of the various multipliers are:

$$\frac{\partial y}{\partial i_0} = \alpha = \left(\frac{1}{1 - c_y + c_y t_y + \frac{1}{\rho^r} z_{cy}(1 - t_y)} \right) > 0 \tag{20}$$

$$\frac{\partial y}{\partial g_0} = \alpha = \left(\frac{1}{1 - c_y + c_y t_y + \frac{1}{\rho^r} z_{cy}(1 - t_y)} \right) > 0 \tag{21}$$

$$\frac{\partial y}{\partial t_0} = (-c_y + \frac{1}{\rho^r} z_{cy})\alpha = \left(\frac{-c_y + \frac{1}{\rho^r} z_{cy}}{1 - c_y + c_y t_y + \frac{1}{\rho^r} z_{cy}(1 - t_y)} \right) < 0 \tag{22}$$

Note that these investment and fiscal multipliers are deceptive and even erroneous for a monetary economy. They are based on the commodity market, which is a very limited part of the economy, and ignore the monetary sector, which is also needed for determining aggregate demand in the economy. Further, they ignore the supply side of the economy. The appropriate multipliers for the economy are those derived after a general analysis of all the sectors of the whole economy. Therefore, the usefulness of the above multipliers lies only in studying the determinants of the shifts of the IS curve.

13.4 The monetary sector: determining the appropriate operating target of monetary policy

The specification of the monetary sector for any given economy requires the prior determination of the primary monetary instrument used by its central bank. Benjamin Friedman (1990) provides an extensive discussion of the various instruments and targets of

monetary policy, and the deviation of the appropriate instrument from an assumed target. The two main instruments available to the central bank are the money supply (often the monetary base) and the interest rate. The specification of which of these is the primary instrument can be determined from knowledge of the central bank's operations, from its statements, or from causality tests.

While some central banks are very forthcoming or transparent in their statements as to which is their operating target, this is not so for every central bank. In the latter case, it becomes necessary to perform a causality test between the money supply or, preferably, the monetary base, and the relevant interest rate. The relevant interest rate would be the central bank's discount rate for loans to commercial banks or the interest rate, such as the overnight loan rate directly controlled by it. Causality could be ascertained by the Granger-causality test (see Chapter 7).

In the face of continuing shifts in the money demand function due to financial innovations, central banks in several of the industrialized countries, including the USA, Canada and the UK, have clearly announced their use of the interest rate as the primary monetary policy instrument. However, this is not so for many other countries. For instance, in many developing countries with a large informal financial sector, the central bank may rely on the money supply as the instrument that provides better control over the economy. Given the possibility of variations across countries, this chapter presents two alternative models of the money market. The one presented first takes the money supply as being exogenously set by the central bank and leads to the IS–LM model of aggregate demand. This is followed by the second model, which assumes that the central bank sets the interest rate as the exogenous monetary policy variable. This model dispenses with the LM curve (Romer, 2000) and leads to the IS–IRT model of aggregate demand. Both models are specified for the open economy.

13.5 Derivation of the LM equation

This section deals with the case where central bank chooses the money supply as its monetary policy instrument and assumes that the money supply is exogenously determined by the central bank.

Money demand

Earlier chapters on the demand for money established that this demand should be studied in real rather than nominal terms and that, in its simplest form, the demand for real balances m^{d} is a function of real income y and the nominal interest rate r. Assuming a linear relationship for simplification, the demand for real balances is specified by:

$$m^{\mathrm{d}} = m^{\mathrm{d}}(y, R, FW_0) = m_y y + (FW_0 - m_R R) \tag{23}$$

where:

m	=	real money balances
$m_y y$	=	real transactions balances
$(FW_0 - m_R R)$	=	speculative/portfolio demand for real balances
$m_R R$	=	portfolio demand for bonds
R	=	nominal interest rate
FW_0	=	real financial wealth

The dependence of money demand on the rate of interest arises from several sources, such as the transactions demand (Chapter 4), the speculative demand (Chapter 5) and the precautionary and buffer stock demand (Chapter 6). However, for historical reasons related to Keynes's treatment of money demand, the term $(FW_0 - m_R R)$ is often referred to as the speculative demand for money.[14] We will mostly refer to it as the interest-sensitive money demand.

The examination of the money supply process showed that the supply function for money should be specified in nominal terms and that, if the central bank allowed the economy to determine it, it would include the interest rate and real income among its arguments. Alternatively, the central bank could set its level exogenously and achieve it through its control over the reserve base. In line with the latter assumption, we simplify the money supply function to:

$$M^{\mathrm{s}} = M \tag{24}$$

where M is the exogenously determined money stock.

Equilibrium in the monetary sector requires that

$$M = Pm^{\mathrm{d}} \tag{25}$$

Hence, the equilibrium condition in terms of real balances is:

$$M/P = m_y y + (FW_0 - m_R R) \tag{26}$$

Equation (26) specifies those combinations of y and R that maintain equilibrium in the monetary market. It is called the LM relationship, where L stands for liquidity preference (i.e. the demand for money) and M stands for the money supply. Note that while the nominal stock of money is exogenously given and in the control of the central bank, the real value of that stock, M^{d}/P, depends upon the equilibrium price level P in the economy and incorporates choices made by the public, including those on its demand for money.

The LM relationship can also be rewritten as:

$$y = \frac{1}{m_y}\left[\frac{M}{P}\right] + \frac{m_R}{m_y}R - \frac{FW_0}{m_y} \tag{27}$$

where $1/m_y$ is the real balances multiplier; an increase in real balances by one unit increases real income by $1/m_y$.

Equation (27) is plotted in Figure 13.1 as the LM curve. Since $\partial y/\partial R > 0$ in (27), the LM curve has a positive slope. The intuitive explanation for this slope is: an increase in y along the horizontal axis increases the transactions component of the demand for money, requiring the public to reduce its speculative component $m_R R$, which the public will do only at a higher rate of interest. That is, an increase in real income must be accompanied by an increase in the rate of interest for equilibrium to be preserved in the monetary sector.

14 Note that the money demand function is often simplified to $m^{\mathrm{d}} = m_y y - m_R R$. However, as noted in the text, whether the money demand function is written in this form or as in (23), the interest-sensitive element can also arise from the transactions demand analysis presented in Chapter 4. This is especially important since Chapter 5 has argued that the modern economy with a variety of riskless assets might not have a positive demand for speculative balances.

13.5.1 The link between the IS and LM equations: the Fisher equation on interest rates

The Fisher equation for interest rates asserts that in perfect capital markets, the nominal and real interest rates are related by the equation:

$$(1+R)=(1+r)(1+\pi^{e}) \tag{28}$$

where R is the nominal interest rate and r the real one. π^{e} is the expected rate of inflation. Assuming low values of r^{e} and π^{e}, $r^{e}\pi^{e}\to 0$, so that the preceding equation becomes:

$$r^{e}=R-\pi^{e}$$

That is, the real rate that the investor expects to receive equals the nominal rate minus the expected loss of the purchasing power of money balances through inflation. Comparative static analysis with a given money supply does not include the determination of π^{e} so that, for simplification, it can be set at zero or treated as being exogenously determined. Therefore, for simplification, the following analysis assumes that $R=r$. Alternatively, if inflation did occur and needed to be incorporated into the money demand function, we could use the simple form of the Fisher equation, $R=r+\pi^{e}$, to write the money demand function as:

$$m^{d}=m^{d}(y,R)=m_{y}y+(FW_{0}-m_{R}(r+\pi^{e}))$$

In this case, the LM equation would become:

$$y=\frac{1}{m_{y}}\left[\frac{M}{P}\right]+\frac{m_{R}}{m_{y}}(r+\pi^{e})-\frac{FW_{0}}{m_{y}} \tag{27$'$}$$

We choose to base the further development of our model on (27) rather than on (27′), thereby implicitly assuming a comparatively static, inflation-free environment.

13.6 Aggregate demand for commodities in the IS–LM model

If one focuses on the expenditure and monetary sectors only, the IS and LM curves in the (y,r) space convey the impression that they are enough to determine real income and the rate of interest. This is again incorrect. What is being determined by these markets and curves is merely the demand for commodities. This demand has to be set against the supply of commodities to determine actual real income or output in the economy and the price level at which this output will be traded. To proceed along this route, we have to first derive from the IS–LM analysis the demand for output as a function of the price level.

To derive the demand y^{d} for commodities as a function of the price level P, combine the IS and LM equations (17) and (27): the procedure is to solve the LM equation (27) for r (or R) in terms of y and P, and substitute this value of r in the IS equation (17). The resulting equation is:

$$\begin{aligned} y^{d}=&\left(\frac{1}{1-c_{y}+c_{y}t_{y}+\frac{1}{\rho^{r}}z_{cy}(1-t_{y})+i_{r}\frac{m_{y}}{m_{R}}}\right)\cdot\left(\left\{c_{0}-c_{y}t_{0}+i_{0}-\frac{i_{r}}{m_{R}}FW_{0}\right.\right.\\ &\left.\left.+\frac{i_{r}}{m_{r}}\frac{M}{P}+g_{0}+x_{c0}-x_{c\rho}\rho^{r}\right\}+\frac{1}{\rho^{r}}\left\{-z_{c0}+z_{cy}t_{0}-z_{c\rho}\rho^{r}\right\}\right) \end{aligned} \tag{29}$$

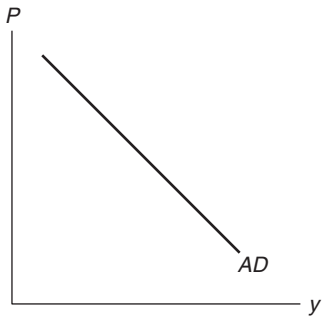


Figure 13.2

Equation (29) specifies the aggregate demand for commodities, which is inversely related to the price level P. (29) specifies those combinations of (y, P) that simultaneously maintain equilibrium in the expenditure and monetary sectors. Figure 13.2 plots equation (29) in the (y, P) space. Designate this curve as the AD (aggregate demand) curve. It has a negative slope.

Note that our demand function is for an open economy. Specifying $\rho^{\mathrm{r}} = \rho \cdot P/P^{\mathrm{F}}$ and assuming a flexible exchange rate would make the nominal exchange rate ρ endogenous by virtue of the balance of payments equilibrium condition, so that it will need to be replaced by its function, which does not have a generally accepted form in macroeconomics. The resulting form of the IS equation would become more complex and would also be one that would be generally less acceptable than (29). Therefore, we do not proceed with this substitution but proceed with (29), while keeping in mind that $\rho^{\mathrm{r}} = \rho \cdot P/P^{\mathrm{F}}$.

Investment and fiscal multipliers for aggregate demand

The investment and government expenditures *multipliers for real aggregate demand* now become:

$$\frac{\partial y^{\mathrm{d}}}{\partial i_0} = \frac{\partial y^{\mathrm{d}}}{\partial g_0} = \left(\frac{1}{1 - c_y + c_y t_y + \frac{1}{\rho^{\mathrm{r}}} z_{cy}(1 - t_y) + i_r \frac{m_y}{m_R}} \right)$$

$$= \left(\frac{m_R}{m_R - m_R c_y + m_R c_y t_y + \frac{1}{\rho^{\mathrm{r}}} m_R z_{cy}(1 - t_y) + i_r m_y} \right) \tag{30}$$

These are smaller than those given by (20) and (21) for the expenditure sector alone. The intuitive explanation for this change is that a given increase in autonomous investment i_0 increases income through the multiplier, which increases the transactions demand for money. The decreased supply of money balances left for speculative purposes increases interest rates, which forces a reduction in induced investment $i_r r$. Total investment, therefore, rises by less than the initial increase in autonomous investment. Note that this multiplier becomes zero if there are no money balances that can be released from speculative to transactions purposes, as would be the case if m_R were zero. Fiscal policy would then have no effect on aggregate demand.

The effectiveness of fiscal policy in increasing aggregate demand thus depends not only upon the parameters of the expenditure sector but also upon those of the monetary sector and the foreign trade sector. In particular, fiscal policy cannot affect aggregate demand when $m_R = 0$; that is, the demand for money – just as its supply – becomes insensitive to the rate of interest.

Money multiplier for aggregate demand

The money supply multiplier, i.e. the increase in aggregate demand for an increase in the nominal money supply, is given by:

$$\frac{\partial y^{\mathrm{d}}}{\partial M} = \left(\frac{1}{1 - c_y + c_y t_y + \frac{1}{\rho^{\mathrm{r}}} z_{cy}(1 - t_y) + i_r \frac{m_y}{m_R}} \right) \frac{i_r}{m_R} \frac{1}{P}$$

$$= \left[\frac{i_r}{m_R P - m_R P c_y + m_R P c_y t_y + \frac{1}{\rho^{\mathrm{r}}} m_R P z_{cy}(1 - t_y) + P i_r m_y} \right] \tag{31}$$

The money multiplier depends upon the relative magnitudes of i_r (the interest sensitivity of investment) and m_R (the interest sensitivity of money demand). If $m_R = 0$, the money multiplier is simply $1/m_y$ and if $i_r = 0$, the real balances multiplier is zero. That is, changes in the money supply cannot affect aggregate demand if investment is insensitive to the rate of interest, i.e. $i_r = 0$, or if the interest rate cannot be changed by monetary policy, which occurs in the liquidity trap, i.e. if $m_R \to \infty$. In the latter case, any increase in the money supply is absorbed by the infinitely elastic demand for money at the existing interest rate. This case of an infinitely elastic demand for money at the existing interest rate is that of the *liquidity trap*. Following Keynes (1936) and the arguments on this concept in Chapter 2 of this book, we take the view that, while the liquidity trap is of intellectual curiosity from an analytical standpoint as a limiting case,[15] it has no practical relevance and can be ignored for analyses of the actual functioning of economies.[16]

Outside the unlikely limits of zero for the money multiplier, the actual magnitude of this multiplier is of little significance since, given the size of the multiplier, any desired change in aggregate demand can be produced by an appropriate change in the money supply, where there is virtually zero difference in cost between changes of different magnitudes in the money supply; i.e. within the usual range of operations, a somewhat larger open market operation does not cost more than a smaller one.

15 This would be so if the interest rate on bonds were zero. Many economists claim that this was roughly so in Japan at times in the last few decades. However, with the return on bonds being zero, the public, with excess money balances, is likely to invest the excess in real assets, such as housing, or in domestic and foreign stocks, etc., so that, while there may be a liquidity trap *vis-à-vis* bonds, it is not likely to occur if all assets are incorporated into the analysis – unless all of them have an expected return equal to zero, which rarely, if ever, happens.

16 The earlier treatment of the liquidity trap was in Chapter 2. We note here again the assertion in Keynes (1936), even though his book was written in the midst of the Great Depression, that he did not consider the liquidity trap to have ever existed in practice.

13.6.1 Keynesian–neoclassical synthesis on aggregate demand in the IS–LM model

Empirical findings and analytical development during the 1950s and early 1960s on the interest elasticities of investment and money demand led to the recognition that, for normally functioning industrialized economies, neither $i_r = 0$ nor $m_R = 0$. Further, such an economy does not have $m_R \to \infty$. Consequently, these extreme cases could be discarded for the relevant macroeconomic analysis, so that the analysis became confined to $i_r < 0, 0 < m_R < \infty$. Acceptance of these results resulted in what came to be known as the "*Keynesian–neoclassical synthesis*" on the determination of aggregate demand. In this synthesis, as is obvious from the preceding aggregate demand equation, both monetary and fiscal policies can change aggregate demand in the economy. Note that this assertion is one about aggregate demand, which equals nominal national income, and not about output, which equals real national income.

The Keynesian–neoclassical synthesis was upset by Barro's (1974) contribution of the Ricardian equivalence theorem, though the majority of the profession continues to believe in the above Keynesian–neoclassical determination of aggregate demand rather than in the form resulting from its incorporation of Barro's Ricardian equivalence theorem.

13.7 Ricardian equivalence[17] and the impact of fiscal policy on aggregate demand in the IS–LM model

While the theory of Ricardian equivalence is not directly relevant to the impact of monetary policy, it is relevant to the relative impact of monetary versus fiscal policy on nominal income in the economy. Ricardian equivalence is the proposition that (bond-financed) fiscal deficits merely postpone taxation without changing the economy's aggregate demand in the current period, as well as in the future. In terms of the IS–LM diagram, Ricardian equivalence implies that fiscal deficits and surpluses do not shift the IS curve from its position when the budget is balanced. This runs counter to the analysis of aggregate demand presented so far in this chapter and to its results on the effects of fiscal policy.

Ricardian equivalence rests on the propositions that, in perfect capital markets,

1 The future tax liability[18] imposed by a bond-financed deficit equals the amount of the current deficit.[19] This liability arises from the future payment of the interest rate on the bonds and the repayment of the principal amount borrowed.
2 The government supplies exactly the goods that the consumers would have bought on their own, so that there is perfect substitutability – in both consumption and production – between government-supplied goods and privately demanded goods.[20] Consequently, the arguments of the consumer's utility function are not the privately bought goods but the sum of the privately bought and the government-provided goods.

17 Although David Ricardo had raised the main idea behind Ricardian equivalence, he rejected its empirical significance. The modern version of this concept is due to Barro's contributions (Barro, 1974).

18 The tax liability imposed by a bond is defined as the present discounted value of the future interest payments and eventually of the principal at maturity, as specified by the bond.

19 The appendix to this chapter establishes this.

20 This assumption is usually not transparent in the standard presentations of Ricardian equivalence since they normally derive the effects of a deficit created through a reduction in taxes. However, deficits do often come into being through an increase in government expenditures. The more general analysis of deficits to cover both possibilities requires the above assumption.

3 Consumers are assumed to have infinitely long lives or to possess an intergenerational utility function encompassing the utility from their own and their descendants' consumption. This intertemporal utility function can be stated as:

$$U(.) = U((c^{\mathrm{p}} + c^{\mathrm{g}})_1, (c^{\mathrm{p}} + c^{\mathrm{g}})_2, \ldots, (c^{\mathrm{p}} + c^{\mathrm{g}})_n) \tag{32}$$

where n can be made arbitrarily large to encompass the period by which the debt used to finance the current deficit is retired and where c^{p} is privately bought consumer goods and c^{g} is government-provided consumer goods.

4 The individual maximizes the intertemporal utility function, given the assumptions of perfect foresight and perfect capital markets, subject to his intertemporal budget constraint. This intertemporal personal budget constraint has the form:

$$\Sigma_t \Psi_t {c^{\mathrm{p}}}_t = \Sigma_t \Psi_t (y_t - T_t) \quad t = 1, \ldots, n \tag{33}$$

where T is tax payments/revenues and Ψ_t is the discount factor $(= 1/(1-r)^{t-1})$ for period t. (33) is the usual statement that the present discounted value of the privately bought goods equals the present value of disposable income.

5 The government's intertemporal budget constraint, under the assumption of a no ponzi scheme[21] is:

$$\Sigma_t \Psi_t {c^{\mathrm{g}}}_t = \Sigma_t \Psi_t T_t \quad t = 1, \ldots, n \tag{34}$$

Equation (34) assumes that there is no outstanding debt inherited at the beginning of period 1 and asserts that the government-provided goods have to be paid for through taxes over time. This constraint rules out the monetary financing of deficits and indirectly incorporates their financing by a new issue of bonds. The implicit assumption that the discount factors are identical for the private sector in (33) and the government in (34) requires perfect, unregulated capital markets – as well as identical risk premiums for the private sector and the government. Defining the deficit in period t as $(c^{\mathrm{g}} - T)_t$, since we have not imposed the requirement that $T_i = {c^{\mathrm{g}}}_i$, for all i, the government can run a deficit or surplus, or neither, in any particular period i. If there is a deficit in period i, with ${c^{\mathrm{g}}}_i > T_i$, it is covered by issuing bonds at the market rate of interest, so that higher taxes will have to be paid in the future to cover the interest on the bonds and their redemption in the future.[22]

Equations (33) and (34) imply that:

$$\Sigma_t \Psi_t (c^{\mathrm{p}} + c^{\mathrm{g}})_t = \Sigma_t \Psi_t y_t \tag{35}$$

which does not include government surpluses and deficits as a variable. Note that y is total income, not disposable income.

21 A ponzi scheme is one where the borrower does not expect to eventually pay back his loan.

22 Given the public's resentment over high tax rates and the commonly recognized dependence of labor supply and tax evasion on the marginal tax rate, we need to recognize – and evaluate the validity of – an implicit assumption on the labor supply function in establishing the Ricardian equivalence hypothesis. This assumption is that the labor supply over time is determined by the intertemporal maximization, so that the future labor supply is not a function of the future tax rate. Therefore, high current deficits resulting in high future tax rates do not reduce the future's labor supply.

The public's choice can now be formulated as: find the total consumption $(c^{\mathrm{p}}{}_t + c^{\mathrm{g}}{}_t)$ that maximizes (32) subject to (35). This will yield the optimal consumption pattern as:

$$(c^{\mathrm{p}} + c^{\mathrm{g}})_t{}^* = f(\Psi, \mathrm{PDW}) \quad t = 1, 2, \ldots, n \tag{36}$$

where:

$$\mathrm{PDW} = \Sigma_t \Psi_t y_t \tag{37}$$

and $\boldsymbol{\Psi}$ is the vector of discount factors. Note that PDW is the present discounted value of total income y (rather than of disposable income) so that the optimal total consumption levels do not include tax revenues, tax rates or government expenditures among their determinants. Assuming an interior solution and the independence of the present discounted value of wealth from fiscal variables, if there is an increase (decrease) in the exogenous supply of government-provided goods in t, the private sector will reduce (increase) its own purchases of goods in t by exactly the same amount, leaving $(c^{\mathrm{p}} + c^{\mathrm{g}})_t$ invariant to government-provided goods $c^{\mathrm{g}}{}_{\mathrm{t}}$, tax revenues, deficits and the government debt. That is:

$$\frac{\partial (c^{\mathrm{p}} + c^{\mathrm{g}})_{t+i}}{\partial x_{t+j}} = 0 \quad \text{for all } i, j \tag{38}$$

where x can be government expenditures, tax revenues or deficits. This result is known as Ricardian equivalence.[23]

Ricardian equivalence is usually stated as the proposition that national saving is invariant with respect to the fiscal deficit. National saving s^n, by definition, is:

$$s^n{}_t = y_t - (c^{\mathrm{p}} + c^{\mathrm{g}})_t \tag{39}$$

where $(c^{\mathrm{p}} + c^{\mathrm{g}})_t$ is invariant to the size of fiscal deficits. Further, if it is also assumed that a shift in supply from the private to the public good or vice versa has no impact on the aggregate output y, $\partial y_t / \partial x_t = 0$. This assumption means that the efficiency of production by public and private sector units is identical, which need not always be valid. Therefore:

$$\frac{\partial s^n{}_{t+i}}{\partial x_{t+j}} = 0 \quad \text{for all } i, j \tag{40}$$

Hence, $s^n{}_t$ is invariant with respect to fiscal expenditures and deficits. Further, since national saving is the sum of private and public saving, where the latter equals the government budget

23 We have stated the assumptions and derivation of the Ricardian equivalence theorem somewhat differently from Barro's (1974). In his analysis, the consumer has a time-separable utility function:

$$U(c_t, \ldots) = u(c_t) + \sum_{i=t+1}^{\infty} \beta^{i-t} u(c_t)$$

Such a consumer is sometimes referred to as "dynast." The consumer maximizes this utility function subject to $W_{t-1} = R(W_t - c_t] + y_{t-1} \cdot c_{t+i}$ is the consumption of the $(t+i)$th generation (or the $(t+i)$th period), W is wealth, y is labor income, R is the intergenerational interest rate and β is the personal discount factor.

surplus, private saving must fall (rise) by the amount of an increase in this surplus (deficit). This is another way of stating Ricardian equivalence.

The appendix to this chapter relates two of the common propositions of Ricardian equivalence to the evolution of the public debt.

Monetary and fiscal policies in the macroeconomic model with Ricardian equivalence

To incorporate Ricardian equivalence into the IS-LM(AD) model of this chapter, we need to modify its consumption function to accord with the Ricardian results. Designating $(c^{\mathrm{p}} + c^{\mathrm{g}})_t$ by c'_t and dropping the time subscript t as a simplification for short-run comparative static analysis, the short-run Ricardian consumption function $c'(y)$ will be:

$$c' = c'(y) = c'_0 + c'_y y \tag{41}$$

Note that c' refers to the total of privately bought and government-provided goods, and depends on total income rather than merely on disposable income. Further, there is a change in the meanings of the parameters of the consumption function from those in the neoclassical model presented earlier in this chapter. In (41), $c'_y = \partial(c^{\mathrm{p}} + c^{\mathrm{g}})/\partial y$, so that c'_y is the sum of the private and government marginal propensities to consume. Aggregate expenditures in the open economy after the integration of government-provided goods into the preceding consumption function are now specified by:

$$e = c' + i + (x_c - z_c/\rho^{\mathrm{r}}) \tag{42}$$

which no longer includes government expenditures, since these are included in (36). In equilibrium, $e = y$, so that (41) and (42) modify the open economy IS relationship to:

$$y = e = c' + i + (x_c - z_c/\rho^{\mathrm{r}})$$

$$y = \left(\frac{1}{1 - c'_y + \frac{1}{\rho^{\mathrm{r}}} z_{cy}}\right) \cdot [\{c'_0 + i_0 - i_r r + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} + (1/\rho^{\mathrm{r}}) \cdot \{-z_{c0} - z_{c\rho}\rho^{\mathrm{r}}\}] \tag{43}$$

The aggregate demand function for the open economy becomes:

$$y^{\mathrm{d}} = \left(\frac{1}{1 - c'_y + \frac{1}{\rho^{\mathrm{r}}} z_{cy} + i_r \frac{m_y}{m_R}}\right) \cdot$$

$$\left((c'_0 + i_0 - \frac{i_r}{m_R} FW_0 + \frac{i_r}{m_R}\frac{M}{P} + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}) + \frac{1}{\rho^{\mathrm{r}}}(-z_{c0} - z_{c\rho}\rho^{\mathrm{r}})\right) \tag{44}$$

Neither (43) nor (44) includes any of the fiscal variables, so that the changes in government expenditures, taxes and deficits will not shift the IS curve in the IS–LM Figures 13.1 or the AD curve in Figure 13.2, and therefore will have no effects on the macro economy. That is, the fiscal variables cannot be used to change aggregate demand and nominal income in the economy, so they do not provide a policy tool for macroeconomic stabilization. Note also that, under Ricardian equivalence, fiscal policy does not affect the interest rate.

Since the equations of the monetary sector in the neoclassical model are not affected by Ricardian equivalence, the only policy tool available for changing aggregate demand and nominal income under Ricardian equivalence is monetary policy. Ricardian equivalence, therefore, buttresses the importance of monetary policy in the economy.[24] From (44), Ricardian equivalence changes the money supply multiplier $\partial y^{\mathrm{d}}/\partial M$ to:

$$\begin{aligned}\frac{\partial y^{\mathrm{d}}}{\partial M} &= \left(\frac{1}{1 - c'_y + \frac{1}{\rho^{\mathrm{r}}} z_{cy} + i_r \frac{m_y}{m_R}}\right) \cdot \frac{i_r}{m_R} \frac{1}{P} \\ &= \left(\frac{i_r}{m_R P - m_R P c'_y + \frac{1}{\rho^{\mathrm{r}}} m_R P z_{cy} + P i_r m_y}\right) \end{aligned} \tag{45}$$

Empirical validity of the Ricardian equivalence hypothesis

Ricardian equivalence requires some very strong and seemingly unrealistic assumptions, so that there are serious questions about its validity. Its validity can be judged on the basis of its assumptions, its implications for the impact of fiscal variables, especially deficits, on aggregate demand and its implications for intergenerational saving behavior.

The impact of fiscal variables on aggregate demand can be tested by the St Louis (monetarist) equation. The St Louis school provided in the late 1960s and early 1970s an empirical procedure for estimating the relationship between nominal income and the money supply. This was the estimation of a reduced-form equation of the form:

$$Y_t = \alpha_0 + \Sigma_i a_i M_{t-i} + \Sigma_j b_j G_{t-j} + \Sigma_s c_s Z_{t-s} + \mu_t \tag{46}$$

where:

Y = nominal national income
M = vector of the past and present nominal values of the appropriate monetary aggregate
G = vector of the past and present values of the fiscal deficit
Z = vector of the other independent variables
μ = disturbance term

The appropriate form of the preceding St Louis equation for testing Ricardian equivalence would use nominal income as the dependent variable. Its earliest estimation was by researchers at the Federal Reserve Bank of St Louis (Andersen and Jordan, 1968), who found[25] for the USA that the marginal impact of fiscal policy was positive for the first year and then turned negative, with a multiplier of only about 0.05 over five quarters.[26]

24 This is consistent with the pre-Keynesian emphasis on monetary policy and a general neglect of fiscal deficits as a policy tool.

25 Another of their findings was that the money aggregates had a strong, positive and rapid impact on nominal income and this impact was more significant than that of fiscal policy. The marginal money-income multiplier was about 5 over five quarters.

26 Numerous applications of the St Louis equation showed that the empirical findings from it differed among countries, periods and definitions of the policy variables. However, the researchers' basic conclusion that the money supply changes have a strong short-term impact on the economy remained fairly robust. Their finding of the insignificance of the impact of fiscal policy on aggregate demand was not, in general, confirmed by other studies.

However, subsequent numerous estimations for different countries and different sample periods tended to show that deficits did have a significant positive impact on aggregate demand.

Ricardian equivalence embodies a special form of the intergenerational saving and bequest function, so that its validity can also be tested directly from the data on savings and bequests. It has been rejected by household panel studies on intergenerational saving and bequest behavior. Among studies that reject it is that of Carroll (2000), who shows that Barro's dynastic model of household consumption and intergenerational saving proves to be a poor description of the behavior of the population, especially of its richest members who contribute the largest proportion of saving in the economy.

13.8 IS–LM model under a Taylor-type rule for the money supply

A feedback monetary policy rule often proposed for the interest rate policy pursued by the central bank is the Taylor rule (Taylor, 1993, 1999). If this rule were adapted to the supply of money to the economy, the money supply rule would be specified by:

$$M^{s}{}_{t} = M^{s}{}_{0} - \lambda_{y}(y_{t} - y^{f}{}_{t}) - \lambda_{\pi}(\pi_{t} - \pi^{T}) \qquad \lambda_{y}, \lambda_{\pi} > 0 \tag{47}$$

where $M^{s}{}_{0}$ is determined by the long-run money demand for the given values of $y^{f}{}_{t}$ and π^{T}. Implicit in the use of such a rule by the central bank is its acceptance that changes in the money supply do change aggregate demand and inflation, and quite possibly short-run (real) output in a predictable manner, so that monetary policy can reduce the output gap and inflationary pressures in the economy.

13.9 Short-run macro model under an interest rate operating target

Taylor rule

Many central banks, especially in financially developed economies, nowadays choose to use the interest rate, rather than the money supply, as the primary monetary policy instrument, while leaving the money supply endogenous to the economy. Alvarez *et al*. (2001) summarize the current consensus on monetary policy as:

> The central elements of this consensus [about the conduct of monetary policy] are that the instrument of monetary policy ought to be the short-term interest rate, that policy should be focused on the short-term interest rate, and that inflation can be reduced by increasing the short-term interest rate.
>
> Alvarez *et al*. (2001, p. 219).

Few central banks openly admit to following a specific rule, though several empirical studies have shown that they act as if they do so. Among these, the use of the interest rate as the operating monetary policy instrument is often espoused in the form of a Taylor rule (Taylor, 1993, 1999), which is:

$$r^{T}{}_{t} = r_{0} + \alpha(y_{t} - y^{f}) + \beta(\pi_{t} - \pi^{T}) \qquad \alpha, \beta > 0 \tag{48}$$

where r^{T} is the *real* interest rate target of the central banks for financial markets, y is real output, y^{f} is full-employment output, π is the actual inflation rate, π^{T} is the inflation rate desired by the central bank, and the subscript t refers to period t. π^{T} is called the *target inflation rate*. Similarly, y^{f} is the *target output level*. $(y_t - y^{\mathrm{f}})$ is (minus of) the output gap. The Taylor rule is a *feedback rule* according to which changes in two indicators, inflation and output, of the actual performance of the economy cause the central bank to change its real interest rate target, under this feedback rule, the central bank would increase its target real interest rate if actual output (or the demand for it) were too high or if inflation were too high, relative to their long-run or desired levels. Taylor used $\alpha = 0.5$ and $\beta = 0.5$, without estimating their values. The usual practice now is to specify the Taylor rule with unspecified values for these parameters, and to estimate them for the country and period being studied. Their relative ratio should reflect the country's central bank's responses, over the sample period, to the output gap and the deviation of inflation from its desired level.

Since central banks set the nominal rather than the real interest rate, the Taylor rule is also often written as:

$$R^{\mathrm{T}}{}_t = \pi_t + r_0 + \alpha(y_t - y^{\mathrm{f}}) + \beta(\pi_t - \pi^{\mathrm{T}}) \quad \alpha, \beta > 0 \tag{48′}$$

which specifies the nominal interest rate R set by the central bank.

The objective of the manipulation of the interest rate by the Taylor rule is to engineer inflation and output back to their target levels and to do so through a gradual adjustment pattern. By implication, under the Taylor rule, monetary policy is not used to respond to shocks to the macroeconomy that do not affect the output gap and the deviation of inflation from its desired level, so that this rule limits the goals of the central bank to the output gap and inflation.

Under the Taylor rule, as is clear from (48′), if π rises above π^{T} the increase in the target real rate will require that the nominal rate has to rise more than the inflation rate; if π falls below π^{T}, a cut in the target real rate will require that the nominal rate has to fall more than the inflation rate. Such a policy is sometimes depicted as one of "leaning against the wind.' For any given inflation rate, the greater the value of β, the larger will be the change in the real and nominal rates and the stronger the movement to stabilize the economy at the target inflation rate.

Implicit in the Taylor rule are the following propositions about the usual links between the interest rate, output and inflation:

1 An increase in the real interest rate reduces aggregate demand, which reduces inflation, so that the real interest rate and inflation are negatively related.
2 There is a close positive relationship between output and inflation. Given the structure of most economies, there is usually a low positive inflation rate (say, π^{nairu}) when output is at its full-employment level; inflation rises above π^{nairu} as output rises above its full-employment level and falls below π^{nairu} as output falls below that level.[27]

27 In practice, the price level rarely falls. Rather, deficient demand with output somewhat below the full-employment level leads not to falling prices but to a decrease in the inflation rate (Ball, 1994). The extent of the change in inflation depends on many variables such as the output gap, the expected rate of inflation, changes in the cost of imported commodities, the rate of increase in the money supply, etc.

The superscript "nairu" stands for "non-accelerating inflationary rate of unemployment."

The preceding version of the Taylor rule is a *contemporaneous* one, in which the current interest rate varies with the current output gap and the "inflation gap," with the latter defined as $(\pi_t - \pi^T)$. There are at least three other versions of the Taylor rule in the literature. These include a backward-looking (forward-looking) rule in which the current interest rate is set on the basis of past (future) values of the output and inflation gaps. A fourth version derives the Taylor rule from the optimization of a loss function of the central bank. These versions are spelled out in Chapter 15 on the Keynesian paradigm since its current variant (the new Keynesian model) incorporates the Taylor rule.

Note that in the long run $y_t = y^f$ and $\pi_t = \pi^T$, $r^T{}_t = r_0$, so that r_0 has to equal the long-run real interest rate of the economy. Otherwise, the divergence between the real interest rate set by the central bank and the economy's real interest rate would cause long-run disequilibrium in the financial markets of the economy, with consequences for the markets for commodities and labor, so that neither y^f nor π^T would be achieved.

In terms of empirical evidence, some form[28] of the Taylor rule has often done quite well as the central bank's explicit or implicit reaction function for developed, free market economies (Sims, 2001).[29] To cite just one empirical study, Clarida *et al*. (1998) estimate the monetary policy rules for France, Germany, Italy, Japan, the UK and the USA. Using a forward-looking version of Taylor's rule, they report that the central banks of Germany, Japan and the USA can be implicitly taken to have followed inflation targeting and output stabilization functions.[30]

An issue actively pursued in further research on the Taylor rule has been whether or not asset prices and exchange rates should be included in this rule. The argument in favor of their inclusion is that shifts in them can change aggregate demand. However, some part of these shifts are often the result of changes in output and inflation, so that only the impact of their residual shifts on inflation and output would need to be offset through monetary policy. Doing so yields extended forms of the Taylor rule. Many empirical studies report that using some form of an augmented Taylor rule, such as incorporating changes in wealth or house prices or exchange rates, leads to greater stabilization of the economy. However, none of these extended forms has come into general usage in macroeconomic modeling, so we choose not to present them or incorporate them into the analysis of this chapter. The Taylor rule is re-examined in Chapter 15.

28 However, the same form may not do well for all countries or for a given country for all periods. Further, the values of the coefficients may also shift over time, e.g. when the leadership of the central bank changes.

29 Sims points out that, although simple Taylor rules perform quite well, rules that make the change in the interest rate the dependent variable do better than the original Taylor rule where the dependent variable is the level of the interest rate.

30 Kahn and Parrish (1998) discuss the inflation targeting frameworks of various countries, including Canada, the USA and the UK, and present a table summarizing this information.

Judd and Rudebusch (1998) report that reaction functions of the Taylor rule type summarize well key elements of US monetary policy during the 1970–1997 period. Levin *et al*. (1999, 2001) evaluate the robustness of the estimated policy rules to model uncertainty. Their estimates, from US data for five macroeconomic models, show that a simple version of the inflation and output-targeting rule for the US economy performs quite well. Further, responding to an inflation forecast not longer than a year, the rule does better than for forecasts of inflation further into the future. Wang and Handa (2007) find that the Taylor rule can also be validly applied to China, a developing country.

Integration of the interest rate as the operating monetary target into the macroeconomic model

Our desire in the following analysis is to develop a macroeconomic model that has the simplicity of the IS–LM one and is comparable with it. In the IS–LM model, the central bank holds the money supply constant. The corresponding assumption for the interest rate as the monetary policy instrument is that the central bank holds the interest rate constant. Therefore, for the following analysis, we adopt the assumption that the central bank sets the real interest rate at a fixed level, r^{T}, which we designate as the "target rate," and formulate the macroeconomic model under this assumption. This model can be used to analyze the impact of changes in this interest rate, irrespective of whether the changes are made on a discretionary basis or according to a rule such as the Taylor rule.

The assumption made on monetary policy is that the central bank successfully targets and sets the economy's real interest rate r at r_0. That is, under this simple targeting policy,

$$r = r^{\mathrm{T}}{}_0 \tag{49}$$

Plotting this interest rate in the (r, y) space of the IS diagram, we have a horizontal line at the target real interest rate. This is shown in Figure 13.3a by the "interest rate target curve" labeled IRT.[31]

An alternative to the above assumption of a simple fixed interest rate rule is a simple feedback rule such as:

$$r = r_0 + \lambda_y y^{\mathrm{d}} + \lambda_P P \quad \lambda_y, \lambda_P > 0 \tag{49$'$}$$

(49′) can also be modified to a Taylor-type rule but with targeting of price level, rather than inflation, as in:

$$r^{\mathrm{T}}{}_t = r_0 + \alpha(y_t - y^{\mathrm{f}}) + \beta(P_t - P^{\mathrm{T}}) \quad \alpha, \beta > 0 \tag{49$''$}$$

In (49′) and (49″), an increase in aggregate demand causes the central bank to raise the interest rate, so that r and y^{d} are positively related, giving the IRT curve a positive slope, as shown by the IRT curves in Figure 13.3b. An increase in P would not shift the IRT curve under (49) but would shift those in (49′) and (49″) upward, from IRT_0 to IRT_1, indicating an increase in r at any given level of y^{d}.

Note that, as shown later, under interest rate targeting of whatever type, the money supply becomes endogenous to money demand, which the central bank accommodates by appropriate changes in the monetary base. Further, under the simple interest rate targeting in (49), the LM curve becomes horizontal at the set interest rate because the central bank supplies money perfectly elastically to the economy. However, the horizontal nature of the LM curve in this case does not mean that the economy is in the liquidity trap.

31 Note that the IRT curve replaces the usual LM curve, which is based on the assumption that the central bank targets or sets the money supply. The central bank can target either the interest rate or the money supply, but not both, so that the LM curve is really not appropriate under interest rate targeting.

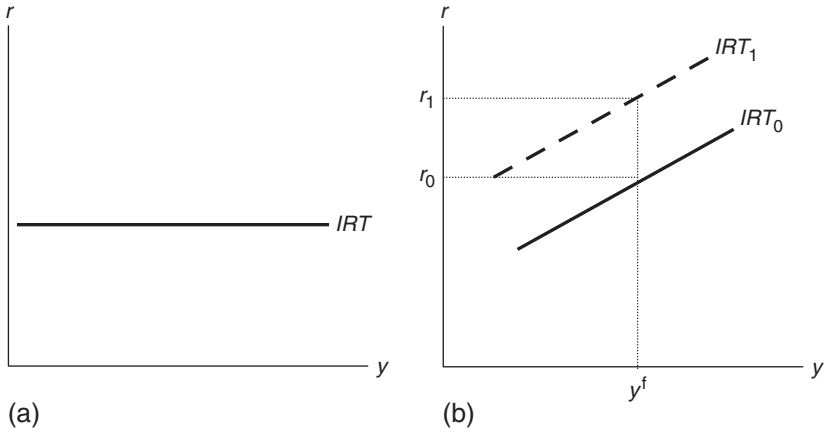


Figure 13.3

Although we have a choice of interest rate rules for monetary policy, we will proceed with its simplest version, which is given in (49). While *some* form of the Taylor rule seems to do quite well empirically for depicting central bank behavior *on average* and *in hindsight*, in practice, the adjustment of interest rates by the central bank at any given time does not happen automatically according to a pre-specified rule and involves considerable discretion and hesitation.[32] Further, central bank preferences on the weights put on the output and inflation gaps tend to shift over time,[33] as they have done in the USA over the past three decades with the changes in the chairman of the Federal Reserve System (Clarida *et al*., 2000). Furthermore, even the assessment of the actual output gap and the current and future course of inflation is usually cloudy, to say the least, and often in dispute.[34] Hence, it seems preferable to present the basic benchmark analysis, intended for general understanding of the macroeconomy, on the general nature of the monetary policy instrument rather than on one of the numerous specific forms of the Taylor rule. We therefore proceed with (49) and leave it to the interested reader to derive the aggregate demand functions under (49′) and (49″) – or one of the other forms of the Taylor rule, some more of which are specified in Chapter 15.

32 Both these were remarkably obvious for the US Fed in 2007 in the midst of the subprime financial crisis in the USA. In fact, John Taylor claimed that the Fed had not followed his rule during 2002 and 2006; had it done so, it would have raised interest rates considerably more than it did, which would have prevented the expansion of mortgages and other loans at very low interest rates to high-risk borrowers, thereby averting the subprime crisis of 2007.

33 In the United States, this shift usually occurs when there is a change in the Chairman of the Fed, as happened in 2006 with the change from Alan Greenspan to Ben Bernanke. But there can also be smaller shifts when the composition of the Federal Open Market Committee changes.

34 For instance, looking at policy formulation in the USA, 2008 was a particularly cloudy year. It started with fairly clear signals of a weakening economy and strong expectations of one or more interest rate cuts. By mid-year, because of rising energy and food prices, concern arose about possibly rising inflation. However, the assessment of the latter was in dispute not only between the Fed Chairman and Wall Street analysts, but, reportedly, also between the Fed policymakers and their own staff economists, and among the members of the Federal Open Market Committee.

13.9.1 Determination of aggregate demand under simple interest rate targeting

The equation for the aggregate demand (AD) for commodities is obtained jointly from the IS equation for the commodity market and the monetary policy equation determining the interest rate. Our earlier analysis implies the IS equation to be:

$$y^{\mathrm{d}} = \alpha[\{c_0 - c_y t_0 + i_0 - i_r r + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} + (1/\rho^{\mathrm{r}}) \cdot \{-z_{c0} + z_{cy} t_0 - z_{c\rho}\rho^{\mathrm{r}}\}] \quad (50)$$

where:

$$\alpha = \left(\frac{1}{1 - c_y + c_y t_y + \frac{1}{\rho^{\mathrm{r}}} z_{cy}(1 - t_y)} \right) > 0 \quad (51)$$

For monetary policy, in order to reduce unnecessary complexity, as discussed above, we assume the simple monetary policy rule:

$$r = r^{\mathrm{T}}{}_0 \quad (52)$$

Substitution of (52) in (50) yields the *AD equation*:

$$y^{\mathrm{d}} = \alpha[\{c_0 - c_y t_0 + i_0 - i_r r^{\mathrm{T}}{}_0 + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} + (1/\rho^{\mathrm{r}}) \cdot \{-z_{c0} + z_{cy} t_0 - z_{c\rho}\rho^{\mathrm{r}}\}] \quad (53)$$

The exogenous policy variables in this model are the fiscal ones of g, t_0 and t_y, while the monetary policy one is $r^{\mathrm{T}}{}_0$. The respective policy multipliers are $\partial y^{\mathrm{d}}/\partial g = \alpha$, $\partial y^{\mathrm{d}}/\partial t_0 = [(-c_y + z_{cy}(1/\rho^{\mathrm{r}}))\alpha]$ and $\partial y^{\mathrm{d}}/\partial r^{\mathrm{T}} = -i_r\alpha$. Hence, both fiscal and monetary policies are effective in increasing aggregate demand. The AD equation is almost the same as the IS equation, so that its investment and fiscal multipliers are identical to the ones derived from the IS equation alone.[35]

Diagrammatic derivation of the AD curve

Aggregate demand in an economy for which the central bank sets the interest rate is given by the intersection of the IS and the IRT curves. In Figure 13.4a, the intersection of these two curves at the given interest rate determines the level of the aggregate demand for domestic commodities. This level is shown as y^{d}. Further, at the given interest rate, a rightward shift of the IS curve because of increases in investment, government expenditures, exports and the other reasons mentioned above, will increase aggregate demand. A cut by the central bank in the interest rate will also increase aggregate demand.[36]

Note that the assumption that monetary policy targets the interest rate rather than the money supply implies that the LM curve is omitted from Figure 13.4a, so that it does not play a direct role in the determination of aggregate demand (Romer, 2000). Consequently,

35 The AD equation and its multipliers would change under a monetary policy rule that makes the interest rate a positive function of y^{d}. Since there is no change in the qualitative results that we use in later analysis, the calculation of the AD equation and its multipliers for the interest rate specified as a positive function of y has been left to the interested reader.

36 These results would also hold under a Taylor rule that makes the interest rate a positive function of y^{d}.

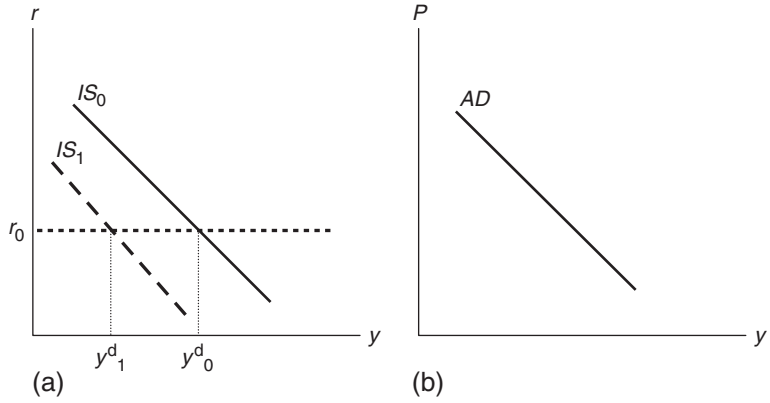


Figure 13.4

the IS–LM diagram specified under an exogenous money supply is replaced under interest rate targeting by the IS–IRT diagram.

We concluded above that, under flexible exchange rates, an increase in P causes a leftward shift of the IS curve: the increase in the real exchange rate due to the increase in the domestic price level makes domestic goods more expensive than foreign ones, so that the substitution effect increases the demand for imported goods and decreases net exports.[37] This leftward shift of the IS curve decreases aggregate demand at the given IRT, as is shown in Figure 13.4a by the shift of the IS curve from IS_0 to IS_1, with a consequent decrease in demand from ${y^{\mathrm{d}}}_0$ to ${y^{\mathrm{d}}}_1$. Conversely, a decrease in P causes a rightward shift of the IS curve, which increases aggregate demand. Therefore, the aggregate demand curve AD in Figure 13.4b is downward sloping.

Note that the IS and AD curves for the open economy incorporate within them the effects of changes in y and r on the nominal exchange rate ρ and, through it, on exports and imports. Note also that an increase in P (or decrease in P^{F}) shifts the IS curve to the left whereas a decrease in P (or increase in P^{F}) shifts it to the right, but changes in P do not shift the AD curve.

13.9.2 Aggregate demand under the Taylor rule

Aggregate demand under simple interest rate targeting was shown to be:

$$y^{\mathrm{d}} = \alpha[\{c_0 - c_y t_0 + i_0 - i_r {r^{\mathrm{T}}}_0 + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} + (1/\rho^{\mathrm{r}}) \cdot \{-z_{c0} + z_{cy} t_0 - z_{c\rho}\rho^{\mathrm{r}}\}] \tag{54}$$

Now, suppose that the central bank follows the Taylor rule,

$${r^{\mathrm{T}}}_t = r_0 + \alpha(y_t - y^{\mathrm{f}}) + \beta(\pi_t - \pi^{\mathrm{T}}) \quad \alpha, \beta > 0 \tag{55}$$

37 This decrease in net exports would cause the nominal exchange rate to depreciate but not by enough to cancel out the initial appreciation of the real exchange rate. Further, under perfect capital mobility, capital flows would not change since the domestic and foreign interest rates would not have changed.

In this case, the AD equation would become:

$$y^{\mathrm{d}} = \alpha[\{c_0 - c_y t_0 + r_0 + \alpha(y_t - y^{\mathrm{f}}) + \beta(\pi_t - \pi^{\mathrm{T}}) - i_r r^{\mathrm{T}}{}_0 + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} + (1/\rho^{\mathrm{r}}) \cdot \{-z_{c0} + z_{cy} t_0 - z_{c\rho}\rho^{\mathrm{r}}\}] \quad (56)$$

which maintains the efficacy of both fiscal and monetary policies in changing aggregate demand; an increase in fiscal deficits increases aggregate demand, while a rise in the interest rate lowers it.

13.9.3 Aggregate demand under the simple interest rate target and Ricardian equivalence

The IS equation under Ricardian equivalence was shown above to be:

$$y = \left(\frac{1}{1 - c'_y + \frac{1}{\rho^{\mathrm{r}}} z_{cy}}\right) \cdot [\{c'_0 + i_0 - i_r r + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} + (1/\rho^{\mathrm{r}}) \cdot \{-z_{c0} - z_{c\rho}\rho^{\mathrm{r}}\}] \quad (57)$$

Therefore, the aggregate demand function with Ricardian equivalence and $r = r^{\mathrm{T}}$ is:

$$y = \left(\frac{1}{1 - c'_y + \frac{1}{\rho^{\mathrm{r}}} z_{cy}}\right) \cdot [\{c'_0 + i_0 - i_r r^{\mathrm{T}} + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} + (1/\rho^{\mathrm{r}}) \cdot \{-z_{c0} - z_{c\rho}\rho^{\mathrm{r}}\}] \quad (58)$$

which implies the ineffectiveness of fiscal policies in changing aggregate demand, while maintaining the impact of monetary policy on aggregate demand.

13.9.4 The potential for disequilibrium in the financial markets under an interest rate target

Demand for money

The real money demand specified earlier for the 1960s Keynesian–neoclassical synthesis is:

$$m^{\mathrm{d}} = m^{\mathrm{d}}(y, R, FW_0) = m_y y + (FW_0 - m_R R) \quad 0 < m_y \leq 1, 0 < m_R < \infty \quad (59)$$

To cover the possibility that y^{d} (aggregate demand) may not equal y^{s} (aggregate output), we have to specify whether y in the money demand equation should be y^{d} or y^{s}. Since money demand depends upon the planned purchases of commodities, which need to be financed through payment in the medium of exchange, money demand would depend on the aggregate demand for commodities rather on aggregate supply. Therefore, the general money demand equation should be more appropriately specified as:

$$m^{\mathrm{d}} = m^{\mathrm{d}}(y^{\mathrm{d}}, R) = m_y y^{\mathrm{d}} + (FW_0 - m_R R) \quad 0 < m_y \leq 1, 0 < m_R < \infty \quad (60)$$

For perfect capital markets, the Fisher equation on interest rates is:

$$R = r + \pi^{\mathrm{e}} \quad (61)$$

where r is set at r^{T} by the central bank and y^{d} is determined by the AD equation (52) as:

$$y^{\mathrm{d}} = \alpha[\{c_0 - c_y t_0 + i_0 - i_r r^{\mathrm{T}}{}_0 + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} + (1/\rho^{\mathrm{r}}) \cdot \{-z_{c0} + z_{cy} t_0 - z_{c\rho}\rho^{\mathrm{r}}\}] \tag{62}$$

Substituting these values of r^{T} and y^{d} in the money demand equation, we get:

$$\begin{aligned} m^{\mathrm{d}} = {} & m_y \alpha[\{c_0 - c_y t_0 + i_0 - i_r r^{\mathrm{T}}{}_0 + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} \\ & + (1/\rho^{\mathrm{r}}) \cdot \{-z_{c0} + z_{cy} t_0 - z_{c\rho}\rho^{\mathrm{r}}\}] - m_R r^{\mathrm{T}} - m_r \pi^{\mathrm{e}} + FW_0 \end{aligned} \tag{63}$$

Supply of money under interest rate targeting

The money market equilibrium condition is:

$$\begin{aligned} M^{\mathrm{s}}/P = {} & m_y \alpha[\{c_0 - c_y t_0 + i_0 - i_r r^{\mathrm{T}}{}_0 + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} \\ & + (1/\rho^{\mathrm{r}}) \cdot \{-z_{c0} + z_{cy} t_0 - z_{c\rho}\rho^{\mathrm{r}}\}] - m_R r^{\mathrm{T}} - m_R \pi^{\mathrm{e}} + FW_0 \end{aligned} \tag{64}$$

For given P and π^{e}, this equation determines the real money supply M/P required for equilibrium. However, the money market on its own cannot change P, whose movement depends upon the aggregate demand for commodities relative to their supply. Nor can the money market change M^{s} (which depends upon the monetary base M0 controlled by the central bank) and the monetary base multiplier ($\partial M/\partial \mathrm{M0}$), which depends on the payments system and public behavior. Therefore, there is no equilibrating mechanism in the money market that will adjust M/P to real money demand, so that unless the central bank provides the nominal money supply required for money market equilibrium, there is a strong potential for disequilibrium in this market.

For the study of dynamic adjustments in disequilibrium, designate the value of M/P specified by real money demand as $(M/P)^{\mathrm{d}}$ and that specified by the existing nominal money supply and price level as $(M/P)^{\mathrm{s}}$. Adopt the plausible hypothesis that if $(M/P)^{\mathrm{d}} > (M/P)^{\mathrm{s}}$, the individual members of the public attempt to get additional money balances by selling bonds, which reduces bond prices, thereby putting upward pressure on the market interest rate. Conversely, if $(M/P)^{\mathrm{d}} < (M/P)^{\mathrm{s}}$, the individual members of the public invest their excess money holdings by buying bonds, which raises bond prices, thereby putting downward pressure on the market interest rate.

These movements in the nominal interest rate will produce a deviation of the market-determined interest rate from the target real interest rate plus expected inflation. This discrepancy will sooner or later produce a sufficient difference and incentive for private lenders to charge the market-determined rate rather than the central bank's target rate. To prevent such a discrepancy, the central bank will need to adjust the monetary base to ensure that the nominal money supply is suitably adjusted. For our model, it should be specified by:

$$\begin{aligned} M^{\mathrm{s}*} = {} & P[m_y \alpha\{(c_0 - c_y t_0 + i_0 - i_r r^{\mathrm{T}}{}_0 + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}) \\ & + (1/\rho^{\mathrm{r}}) \cdot (-z_{c0} + z_{cy} t_0 - z_{c\rho}\rho^{\mathrm{r}})\} - m_R r^{\mathrm{T}} - m_R \pi^{\mathrm{T}} + FW_0] \end{aligned} \tag{65}$$

where P is determined by the dynamic adjustments in the commodity market and the term in the [.] specifies real money demand. The star * on M^{s} indicates the endogenous value of

the money supply that the central bank needs to achieve if the bond market is to maintain the nominal interest rate at $(r^{\mathrm{T}} + \pi^{\mathrm{e}})$.

Note that (65) replaces the traditional LM equation, but is not really the LM equation, since its role now is to endogenously determine the money supply rather than to be an element in the determination of the price level or of aggregate demand. We will designate it as the M^{s*} equation, with * standing for money market equilibrium.

The policy problem: providing the money supply needed for money market equilibrium

Two other considerations need to be added to (65) to determine the money supply that the central bank needs to ensure if disequilibrium in the money market at the set interest rate is to be avoided. One, the functions for the variables in (65) will have stochastic terms. Or, alternatively stated, the demand functions for both commodities and money will be stochastic. Two, if there are lags in the system, the appropriate values of the demand for commodities and for money will be the expected, rather than the current actual values, for the period when the money supply will impact on the economy. These values will involve forecasting errors. In short, it will be difficult for the central bank to predict and provide the money supply needed to support its set interest rate in the money market.

If the central bank were to achieve its target inflation rate π^{T} and support its set interest rate, our arguments imply that the money supply would have to be:

$$M^{s*} = P[m_y\alpha[\{c_0 - c_y t_0 + i_0 - i_r r^{\mathrm{T}}{}_0 + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} + (1/\rho^{\mathrm{r}}) \cdot \{-z_{c0} + z_{cy}t_0 - z_{c\rho}\rho^{\mathrm{r}}\}] - m_R r^{\mathrm{T}} - m_R \pi^{\mathrm{T}} + FW_0] + \mu_t \tag{66}$$

where μ_t is the stochastic component of money demand. If disequilibrium in the financial markets is to be avoided, the actual money supply the central bank would need to provide in period t would differ from that specified by (65) by the actual values of μ_t. The central bank can provide this money supply through its control of the monetary base exercised through open market operations and through changes in reserve requirements or/and by allowing banks to borrow from it.

The case for allowing borrowing by commercial banks from the central bank

The preceding analysis implies that a central bank that chooses to use the interest rate as its primary monetary policy instrument cannot just ignore the money supply.[38] Rather, it has to ensure an appropriate money supply for equilibrium in the money market. Further, money demand may be unstable because of the instability of the IS equation and/or of the money demand equation, as has been the case in recent decades. Furthermore, even if money demand is stable, there is likely to be a stochastic element to it. In addition, the central bank does not directly control the money supply. Its operative instrument is the monetary base, and the multiplier from the monetary base to the money supply is also likely to be stochastic. These factors individually and collectively cause unpredictability of the appropriate money supply

38 This implies that, even under an interest rate targeting policy, the central bank needs to monitor the monetary aggregates in the economy. For another view of the benefits from doing so, see Woodford (2007).

and monetary base. If the central bank is not able to predict and hit its needed amount for equilibrium in the financial markets, it must provide an avenue through which the banking system can, at its own initiative, bring about the required change in the monetary base. Borrowing by commercial banks from the central bank provides such a mechanism and acts as a safety valve against disequilibrium. This role is somewhat distinct from the role of the central bank as lender of last resort, which is to guard against a liquidity crisis for individual banks.

The role of the central bank in accommodating long-term growth in money demand

Since income rises on a long-term basis and inflation increases the price level over time, there is a secular growth in money demand. This secular growth cannot be accommodated through borrowing by the commercial banks from the central bank, since such loans are short-term artifacts. Therefore, unless these borrowings are to be allowed to mushroom over time, as they are never permitted to do, the central bank needs to undertake open market operations at its own initiative to try and achieve the required monetary base.

Further discussion of changes in the money supply is to be found in Chapter 11 on central banking.

13.10 Does interest rate targeting make the money supply redundant?

Interest rate targeting, chosen as the operating target of monetary policy, would make the money supply redundant if the money supply did not exert an influence on aggregate demand and output independently of the interest rate. Rudebusch and Svensson (2002) find support for such redundancy of the money supply. They report, for US data from 1961 to 1996, that while changes in interest rates have a significant effect on output, changes in money do not have an independent *additional* influence on output.

However, money supply need not be redundant as an explanatory variable for output, for several reasons. Among these, one possibility is that movements in money supply have a lagged affect on the long-term interest rate, which affects investment and aggregate demand, so that money supply movements contain information for the future course of output. Another explanation is that changes in the money supply affect the amount and cost of credit, to some extent independently of their impact through interest rates, and the latter have an impact on output (see Chapter 16).

Nelson (2002) and Hafer *et al.* (2007) provide evidence for some independent impact of the money supply in explaining output. Hafer and colleagues report, for US data since 1960, a statistically significant impact of money (M2) on output, even after taking account of changes in the real interest rate. Further, movements in both inside and outside money have predictive power for movements in output. They conclude, therefore, that interest rate targeting does not make the money supply redundant in explaining output changes.

Many economies, especially among the LDCs, have fragmented financial markets, with formal and informal financial sectors as well as black (i.e. illegal holdings of) money. In these economies, it is not a priori clear which target of monetary policy would best provide control over aggregate demand. Offhand, the interest rate seems to be the better instrument for controlling the formal financial sector's impact on commodity demand, while the money supply seems to be more effective in controlling the commodity demand fueled by the informal and black money sectors. Since the two instruments have differential impact on aggregate demand and, especially, on sectoral demands, optimal monetary policy may then

pursue the two instruments in LDCs in a less coordinated manner than in financially developed economies. In this case, in terms of causation, Granger-causality tests should show two-way causality between the interest rate and the money supply.

Note that if the money supply does exert an influence on output independent of its impact through interest rates, then the interest rate would not be the only useful indicator of the need for monetary policy, and should not be the only operating target of monetary policy.

13.11 Weaknesses of the IS–LM and IS–IRT analyses of aggregate demand

The IS–LM and IS–IRT models offer alternative models, depending on the operating targets of monetary policy, of the determination of aggregate demand for commodities, not the supply of commodities, output or the price level. The latter are the focus of the next three chapters.

The IS–LM and IS–IRT models suffer from certain weaknesses. Briefly, these are:

1 They do not effectively distinguish between the real interest rate, which determines investment, and the nominal interest rates, which determines the demand for money and bonds. From the Fisher equation, the difference between these rates is the expected inflation rate, the determination of which requires a theory of expectations in addition to a macroeconomic model for determining the inflation rate, which is a dynamic variable. There is no simple way of incorporating these into the IS–LM and IS–IRT models. The simplifying assumption that is then inserted is that the expected inflation rate is exogenous to the model or the model is for comparative static analysis.
2 The IS–LM and IS–IRT models do not recognize more than two distinct financial assets, money and bonds, whereas more than two such assets could enrich the analysis and explain additional aspects of the macroeconomy. Chapter 16 addresses this concern by differentiating credit and loans from bonds.
3 The IS–LM and IS–IRT models are for the short run in which the capital stock, technology and labor force are held fixed. For certain types of analysis, such as for explaining business cycles and growth, one needs to allow these to vary.
4 The behavioral equations embedded in the standard IS, LM and IRT equations should properly be derived from explicit optimization analysis of rational economic agents, with forward-looking expectations. Some of this analysis is presented in the context of the new Keynesian IS equation in Chapter 15.

The agenda of macroeconomics in recent decades seems to have been to replace the IS–LM and IS–IRT models by a dynamic stochastic analysis with optimizing agents, who form forward-looking expectations. However, the assumptions adopted to make such models tractable are such that none has won dominance over the standard IS–LM and IS–IRT models as the benchmark models of aggregate demand. In any case, we need to keep in mind what these models are meant to do: provide qualitative conclusions on the main determinants of aggregate demand in a short-run comparative static, not dynamic, context. Admittedly, this is a severely restricted agenda, but it is still a very important one in building one (aggregate demand) of the pillars of macroeconomics. Note that this agenda does not include the derivation of any qualitative conclusions on the determination of the price level, output or employment, since any such conclusions would also need the determination of aggregate supply and equilibrium.

13.12 Optimal choice of the operating target of monetary policy

The two main operating targets of monetary policy discussed in this chapter have been:

- monetary aggregates;
- interest rates.

Chapter 11 used diagrammatic IS–LM analysis to present the central bank's choice between them if it wants to minimize the fluctuations in aggregate demand, and that chapter should be reviewed as an introduction to the following mathematical one. That analysis and its following mathematical version were adapted from Poole (1970). Fischer (1990) provides a more extensive discussion and review of the relevant analyses relating instruments to targets of monetary policy. Poole had assumed that the price level was constant. While this was quite a common assumption in the 1960s, it is no longer considered to be realistic or common in modern macroeconomic analysis. Further, taking the price level as constant, Poole assumed that the central bank's objective was to minimize the variance of real output. Since the following analysis does not assume a constant price level and the central bank can only affect output through aggregate demand and expenditures, it makes the assumption that the central bank wants to minimize the variance of aggregate demand.

Using the earlier IS–LM equations for the determination of aggregate demand, the general stochastic form of the IS and LM equations for the *closed* economy can be written as:

$$\text{IS:} \quad {y^{\mathrm{d}}}_t = -\alpha_r r_t + \mu_t \tag{67}$$

$$\text{LM:} \quad (M/P)_t = -m_R R_t + m_y {y^{\mathrm{d}}}_t + \eta_t \tag{68}$$

where the *variables are now defined as deviations from their trend values*, so that the constants in the IS–LM equations have been omitted. The definitions of the symbols are otherwise as given earlier in this chapter. μ_t is the stochastic disturbance to expenditures on commodities and η_t is the stochastic disturbance in the money market. μ_t and η_t are assumed to have a zero mean, be serially uncorrelated and also uncorrelated with each other. They are assumed to be unpredictable and the central bank is assumed to make its decisions prior to observing their actual values. The right side of (67) specifies expenditures on commodities, with α_r capturing the interest-sensitivity of these expenditures. The right side of (68) specifies money demand in real terms.

This IS–LM model needs to be supplemented by a relationship between r_t and R_t. This is provided, for an economy with perfect capital markets, by the Fisher equation:

$$R_t = r_t + {\pi^{\mathrm{e}}}_t \tag{69}$$

where ${\pi^{\mathrm{e}}}_t$ is the expected inflation rate for period t. Holding these expectations constant as a simplifying analytical assumption, R_t and r_t can be treated as identical in the IS–LM model. Hence, replacing R_t by r_t in the LM equation, the model becomes:

$$\text{IS:} \quad {y^{\mathrm{d}}}_t = -\alpha_r r_t + \mu_t \tag{70}$$

$$\text{LM:} \quad (M/P)_t = -m_R r_t + m_y {y^{\mathrm{d}}}_t + \eta_t \tag{71}$$

The central bank observes the values of all terms, except those of the shocks, prior to setting its policy instrument, which is either M_t or r_t. Its objective function is to minimize the expected variance of aggregate demand around its trend value, i.e.

$$\text{minimize } E({y^{\text{d}}}_t)^2 \tag{72}$$

Since y has been defined in terms of deviations from its trend, note that the equilibrium value of y^{d} in the absence of shocks ($\mu = \eta = 0$) would be zero, so that its variance arises only by virtue of μ and η being different from zero.

When the money stock is the policy instrument, we need to solve (70) and (71) to derive ${y^{\text{d}}}_t$, which is given by:

$$\text{IS-LM:} \quad y^{\text{d}} = \frac{\alpha_r M/P + m_R\mu - \alpha_r\eta}{\alpha_r m_y + m_R} \tag{73}$$

Since $E\mu = E\eta = 0$, $E(y^{\text{d}}) = \frac{(\alpha_r M/P)}{(\alpha_r m_y + m_R)}$. Hence, noting that the variables in the current model are being defined in terms of deviations from their trend values, targeting M such that $E({y^{\text{d}}}_t) = 0$ requires $M/P = 0$, which yields:

$$y^{\text{d}} = \frac{m_R\mu - \alpha_r\eta}{\alpha_r m_y + m_R} \tag{74}$$

Therefore, under the assumption that μ and η are uncorrelated and the variance of $\alpha_P P$ is zero, the variance of aggregate demand under the money supply instrument is given by:

$$E^{\text{m}}(y^{\text{d}})^2 = \frac{{m_R}^2{\sigma_\mu}^2 + {\alpha_r}^2{\sigma_\eta}^2}{(\alpha_r m_y + m_R)^2} \tag{75}$$

where the superscript m on E indicates monetary targeting.

When the real interest rate is the monetary policy instrument, the IS equation (70) alone needs to be solved for ${y^{\text{d}}}_t$. Under interest rate targeting, setting r such that $E({y^{\text{d}}}_t) = 0$, its variance becomes:

$$E^{\text{r}}(y^{\text{d}})^2 = {\sigma_\mu}^2 \tag{76}$$

where the superscript r on E indicates interest rate targeting. Monetary targeting is preferable to interest rate targeting if (75) is less than (76), and vice versa. The former requires:

$$\frac{{m_R}^2{\sigma_\mu}^2 + {\alpha_r}^2{\sigma_\eta}^2}{(\alpha_r m_y + m_R)^2} < {\sigma_\mu}^2 \tag{77}$$

which simplifies to:

$${\sigma_\eta}^2 < {\sigma_\mu}^2\left({m_y}^2 + \frac{2m_y m_R}{\alpha_r}\right) \tag{78}$$

Hence, if there are only money market shocks but no commodity market shocks (i.e. $\sigma_\mu = 0$), then interest rate targeting is preferable since doing so perfectly stabilizes aggregate demand; the LM equation and its disturbance term become irrelevant to the determination of aggregate demand. But if there are only commodity market shocks but no money market shocks (i.e. $\sigma_\eta = 0$), then monetary targeting is preferable since doing so reduces fluctuations in aggregate demand. In this case, with a given money supply, a positive commodity market shock raises the interest rate, which reduces interest-sensitive expenditures, thereby reducing commodity demand, so that the original shock to demand is partially offset.

In the general case, the choice of the policy instrument will depend on the relative magnitudes of the shocks and the slope of the IS curve (whose slope is $-1/\alpha_r$) relative to that of the LM curve (whose slope is $1/m_r$). If, for simplification, the term in parentheses on the right-hand side of (78) were ignored, monetary targeting would be preferable to interest rate targeting if the stochastic disturbance in the money market were smaller than in the commodity market. For the money market, assuming that the central bank can precisely control the money supply but does not know the money demand because of the instability of this demand, σ_η occurs because of the volatility of money demand. For the commodity market, if we were to assume that the instability of commodity demand arises only because of the instability of investment (though consumption and net exports can also be volatile), σ_μ occurs because of the volatility of investment demand. These assumptions provide the commonly used, but simplified, statement of the preceding condition as: monetary targeting is preferable if investment is more volatile than money demand, but interest rate targeting is preferable if money demand is more volatile than investment.

The preceding analysis ignores several aspects of the economy. Regarding the money supply, the central bank does not directly control the money supply. It controls the monetary base, which provides rather imperfect control over the money supply for most economies. Further, the money supply and money demand functions may be unstable, so that their parameters shift over time, with the shifts being unpredictable at the decision time. Regarding the interest rates, the central bank can set its discount rate and manage the overnight loan rate for reserves, but these need not, depending on the structure of the financial markets, provide precise control over market rates or over their differentials. Further, the forms of the IS and LM equations used in the preceding analysis are fairly simple ones and ignore such elements as expectations and factors, such as the exchange rate and net exports, relevant to the open economy.[39]

Other possible operating targets of monetary policy

In addition to interest rates and monetary aggregate as the operating targets of monetary policy, other targets proposed for this purpose include nominal income (or aggregate demand) (see McCallum, 1985; Ireland, 1998) and the price level (see Barro, 1986; Ireland, 1998). The desirability of these targets can be assessed from their welfare effects in the context of a given welfare objective function of the public or of the central bank.

39 Walsh (2006) explores the effects of variations in the model for determining the optimal policy instrument.

Conclusions

The technique of analysis used in this chapter has been one of grouping the relationships of the commodity and money market into a single AD equation. This procedure resulted in the IS-LM model for use when the central bank follows a simple money supply rule and the IS–IRT model when it follows a simple interest rate rule. Which rule is relevant to a given economy? If the money demand function is stable and predictable and the central bank can control the money supply and the interest rate equally well, it does not matter whether the central bank follows a simple money supply rule or a simple interest rate rule. But it does matter when the money demand function is not stable. In this case, the use of the interest rate as the monetary policy instrument may be preferable.

Ricardian equivalence added to the assumptions of the macroeconomic model implies that fiscal policy cannot change aggregate demand. This is so for both the IS–LM and IS–IRT models. With fiscal policy thus made ineffective, monetary policy acquires even more importance as the instrument for inducing changes in aggregate demand.

Under interest rate targeting, with the interest rate set exogenously by the central bank, the analysis of this chapter implies that the financial sector can be omitted from the determination of aggregate demand and, therefore, from any influence on the economy. However, such an implication is incorrect: to avoid disorderly financial markets, the money supply has to equal the money demand at the set interest rate. Failure to do so will bring about a different interest rate in the financial markets, with the consequence that the central bank will lose control over the economy's interest rates. Another reason for studying the money supply arises if the appropriate macroeconomic model includes bonds and loans as two non-monetary assets, with loans being supplied by the financial sector and with imperfect substitution between bonds and loans. The analysis of this case is provided in Chapter 16.

Appendix

The propositions of Ricardian equivalence and the evolution of the public debt

I. Define the amount of outstanding bonds, each with a real value of one unit of commodities, at the beginning of period i to be b_i, and assume $b_1 = 0$. The bonds issued in period i receive interest in period $i+1$. The evolution of the value of the public debt b_{i+1} is specified by:

$$b_{i+1} = \sum_{t=1}^{i} \phi_t (c^{g}_{i-t+1} - T_{i-t+1}) \tag{79}$$

where ϕ_t is the interest compounding factor $(= (1+r)^t)$, which multiplies the past deficits to arrive at their present value, and r is the real interest rate. (79) specifies the evolution of the public debt over time under the assumption that there is no default.

To relate Ricardian equivalence to the public debt, we need to incorporate its assumption that any bonds issued by the government to finance current deficits are redeemed by some future date n which is within the representative individual's horizon[40]

40 This horizon can be farther than the representative individual's expected date of death and will cover the lifetimes of his intended beneficiaries.

for consumption planning. Since all outstanding government bonds are redeemed at the end of period n, $b_{n+1} = 0$. Hence, setting $i = n$ in (79):

$$b_{n+1} = \sum_{t=1}^{n} \phi_t (c^{g}_{n-t+1} - T_{n-t+1}) = 0 \tag{80}$$

so that:

$$\sum_{t=1}^{n} \phi_t c^{g}_{n-t+1} = \sum_{t=1}^{n} \phi_t T_{n-t+1} \tag{81}$$

where $\phi_t = (1+r)^t$. Multiplying both sides of (81) by $(1+r)^{-n}$, we get:

$$\sum_{t=1}^{n} \frac{1}{(1+r)^{n-t}} c^{g}_{n-t+1} = \sum_{t=1}^{n} \frac{1}{(1+r)^{n-t}} T_{n-t+1} \tag{82}$$

which is the same as:[41]

$$\sum_{t=1}^{n} \frac{1}{(1+r)^{t-1}} c^{g}_{t} = \sum_{t=1}^{n} \frac{1}{(1+r)^{t-1}} T_t \tag{83}$$

Equation (83) is identical to (34) for the government's intertemporal budget constraint, assumed earlier in this chapter for the derivation of Ricardian equivalence.

II. If the interest rate paid on bonds and the discount rate are identical, Ricardian equivalence incorporates the assertion that the present value of the future tax liability[42] imposed by the bonds issued to finance the deficit equals the deficit itself. To show this, let the deficit d be financed by the current issue of bonds b, so that $b = d$, and designate by PV_b the present value of the future tax liability. With the payment of the interest Rb at the end of each period and the repayment of the principal P $(= b)$ at the end of period n, PV_b is given by:

$$\begin{aligned} PV_b &= \sum_{t=1}^{n} \frac{RP}{(1+R)^t} + \frac{P}{(1+R)^n} \\ &= P\left(\sum_{t=1}^{n} \frac{R}{(1+R)^t} + \frac{1}{(1+R)^n} \right) \\ &= P(1) = P \end{aligned} \tag{84}$$[43]

where R is the nominal interest rate. Since $P = b = d$, $PV_b = d$, so that the present value of the future tax liabilities from a bond-financed deficit equals the deficit itself.

41 To see this, write out the preceding equation in its long form (without the summation sign).

42 That is, the interest payments prior to redemption and the payment of the principal on the redemption of the bonds.

43 If we let $D = 1/(1+R)$, the proof uses the mathematical formula:

$$\sum_{t=1}^{n} D^t = \sum_{t=0}^{n} D^t - 1 = \frac{1 - D^{n+1}}{1 - D} - 1$$

Summary of critical conclusions

- Under certainty, or with a stable and predictable money demand function, it does not matter whether the central bank chooses the money supply or the interest rate as its operating monetary policy target. Under uncertainty and stochastic functions, it does matter.
- For an economy for which it can be validly assumed that the money supply is the monetary policy instrument and is exogenously determined by the central bank, the appropriate model of aggregate demand is the IS–LM model.
- For an economy for which it can be validly assumed that the interest rate is the monetary policy instrument and is exogenously determined by the central bank, the appropriate model of aggregate demand is the IS–IRT model.
- The addition of Ricardian equivalence to both models makes aggregate demand invariant to fiscal policy, but does not qualitatively affect the scope and impact of monetary policy.

Review and discussion questions

1. For a small open economy with a floating exchange rate whose movements continuously ensure equilibrium in the balance of payments, draw the locus of points in the IS–LM diagram (with the interest rate on the vertical axis and real income on the horizontal one) at which equilibrium exists in the foreign exchange market. In this diagram, do we also need to draw a BP curve (designating the balance of payments equilibrium points) in addition to the IS and LM curves for aggregate demand analysis? Discuss why or why not.
2. It is often asserted that there is a national income *identity*, so that in a closed economy without a government sector there is an identity between saving and investment. Explore the implications of this statement by doing the following:

   Given a closed economy without a government sector, modify the IS–LM model by the assumption that saving and investment are identical (i.e. equal under all circumstances and not merely in equilibrium) and answer the following: What is the level of aggregate demand? Is the LM relationship required for its determination? Is any interest rate consistent with equilibrium in the commodity market?

   Now discuss the following statements: "For purposes of a satisfactory theory of money supply and prices, the IS relationship must never be treated as an identity, so that there is only a national income equilibrium condition." "A satisfactory macroeconomic model must incorporate the potential for disequilibrium between saving and investment, even in a closed economy without a government sector."
3. Suppose that the government wants to increase its expenditure g and has the options of financing it by higher taxes, bond issues or increases in the monetary base. Further, when g is increased through bond or monetary financing, the central bank can also undertake offsetting open market operations. What combinations, if any, of financing methods and open market operations will allow the following goals to be met in the IS–LM model:

   (i) no change in aggregate demand;
   (ii) no change in investment;
   (iii) no change in aggregate demand and investment?

4. Present the analysis of the statement that the effects of government deficits on aggregate demand depend on the way in which the deficit is financed.
   Also, analyze the above statement for the following cases:

   (i) the central bank is known to follow the rule of stabilizing the growth of the money supply;
   (ii) government debt (bonds) is only sold to the central bank;
   (iii) the central bank is known to follow the rule of stabilizing the nominal rate of interest.

5. The central bank has decided to adopt one of the following money supply rules:

   (a) $M^{\mathrm{s}} = kPy$
   (b) $M^{\mathrm{s}} = ky$

   where $k > 0$. Show their implications for aggregate demand in the context of (i) the IS–LM model and (ii) the IS–LM model with a zero speculative demand for money (i.e. $m_R = 0$). Is each of these policies viable?
6. One of the banking innovations in the 1960s was the payment of interest on certain types of demand deposits. Assume that interest is paid on money at the rate R_m, which equals (R-x), where x is exogenously determined in nominal terms by market structures and the cost of servicing deposits.

   (i) Use the Baumol–Tobin transactions demand model to derive the demand function for money.
   (ii) Generalizing the above demand function to $m^{\mathrm{d}}(y, R, x)$, show the behavior of the LM curve for shifts in x and P.
   (iii) What is the effect of an increase in x on aggregate demand and the price level in the IS–LM model?
   (iv) Assuming that both r and x always increase by the expected rate of inflation, carry out (ii) and (iii) again.

7. "In a *closed* economy, if the money stock is held constant by the central bank, an increase in the government deficit does not have either short-run or long-run effects on aggregate demand or the interest rate." Discuss in the context of the IS–LM model and the IS–LM model with Ricardian equivalence.
8. Modify the general IS–LM model of this chapter to the appropriate model for a *closed* economy. Specify the IS, LM and aggregate demand equations. What causes the dependence of aggregate demand on the price level?
9. Assuming that the central bank follows a simple interest rate target rule, modify the general IS–IRT model of this chapter to the appropriate one for a *closed* economy. Specify its IS and aggregate demand equations. Is the price level P a variable in these equations? Plot the IS and AD equations and discuss the reasons for their slopes.
10. Assuming that the central bank follows a simple interest rate target rule, write an essay on the determination of the money supply that would maintain equilibrium in the money and bond markets. Discuss how the central bank can ensure this money supply if the money demand function is unstable and unpredictable.
11. Can monetary and fiscal policies be independent of each other? If not, why is their independence postulated in the IS–LM and IS–IRT analyses? Discuss.

12. The "crowding out" hypothesis is the statement that fiscal deficits reduce investment; in the limit, full crowding out means a reduction in investment equal to the fiscal deficit. Is crowding out fully or partially valid in the IS–LM and IS–IRT models? Can complete crowding out occur in the context of (a) a financially advanced economy, (b) a financially under-developed economy?
13. What are the ways or reasons under which fiscal deficits can crowd out private expenditures in the determination of aggregate demand? In this discussion, do not forget the Ricardian equivalence theorem.
14. In the IS–LM model, discuss how each of the following affect the relative efficacy of monetary and fiscal policies:

    (a) the interest elasticity of money demand;
    (b) the interest elasticity of investment;
    (c) the income elasticity of money demand.

15. "It makes no difference whether government expenditures are financed by taxes or bonds." Specify a theoretical basis for this statement. Provide at least two reasons why it may not hold.
16. Discuss the rationale and validity of the following statements: "It makes no difference whether government expenditures are financed by money or bonds." "A fiscal deficit has a larger impact on aggregate demand if it is financed by money than if it is financed by bonds."
17. An empirical study tested for the Ricardian equivalence theorem by estimating the following equation:

$$\Delta A_t = a_0 + a_1 \Delta B_t + \mu_t$$

    where A is the public's net real financial assets (excluding its holdings of government bonds), B is real public debt and μ is a random term. Does $\hat{a}_1 = 1$ confirm the Ricardian equivalence theorem? US aggregate time series tend to yield $\hat{a}_1 = 0$. What would this imply for the validity of the Ricardian equivalence theorem?

    Formulate and specify at least one other estimation equation for testing the Ricardian equivalence theorem.
18. Assume that the IS and money demand equations each have a disturbance term. Further, assume that the central bank can control the economy's interest rate r and money supply M^{s} except for uncontrollable disturbance terms η_t and ν_t, so that:

$$r = r^{\mathrm{T}} + \eta_t$$
$$M^{\mathrm{s}} = M + \nu_t$$

    Use Poole's analysis to derive the conditions for preferring interest rate targeting over money supply targeting, and vice versa.

    Assume that the central bank chooses to target the interest rate in the preceding equation. Derive the money supply that the central bank needs to ensure for equilibrium in the money market. What would happen if it did not ensure this money supply? Discuss.
19. In the preceding question, suppose the central bank uses open market operations to aim at the money supply. Discuss the case for it to allow borrowing from it by commercial

banks at the latter's initiative. Is the argument for such borrowing different from that usually offered for the lender-of-last-resort function of central banks? Discuss.

References

Alvarez, F., Lucas, Jr. R.E. and Weber, W.E. "Interest rates and inflation." *American Economic Review*, 91, 2001, pp. 219–25.

Anderson, L.C., and Jordan, J.L. "Monetary and fiscal actions: a test of their relative importance in economic stabilization." *Federal Reserve Bank of St. Louis Review*, 1968, pp. 11–24.

Ball, L.B. "What determines the sacrifice ratio?" In N.G. Mankiw, ed., *Monetary Policy*. Chicago: University of Chicago Press, 1994.

Barro, R.J. "Are government bonds net wealth?" *Journal of Political Economy*, 82, 1974, pp. 1095–118.

Barro, R.J. "Recent developments in the theory of rules versus discretion." *Economic Journal*, 96, 1986, pp. 23–7.

Carroll, C.D. "Why do the rich save so much?" Chapter 14 in J.B. Slimrod, ed., *Does Atlas Shrug?* Cambridge, MA: Harvard University Press, 2000.

Clarida, R., Galí, J. and Gertler, M. "Monetary policy rules in practise: some international evidence." *European Economic Review*, 1998, 42, pp. 1033–67.

Clarida, R., Galí, J. and Gertler, M. "Monetary policy rules and macroeconomic stability: evidence and some theory." *Quarterly Journal of Economics*, 115, 2000, pp. 147–80.

Fischer, S. "Rules versus discretion in monetary policy." In B.M. Friedman and F.H. Hahn, eds, *Handbook of Monetary Economics*, vol. II. Amsterdam: North-Holland, 1990, Ch. 21, pp. 1155–84.

Friedman, B.M. "Targets and instruments of monetary policy. "In B.M. Friedman and F.H. Hahn, eds, *Handbook of Monetary Economics*, vol. II. Amsterdam: North-Holland, 1990, Ch. 22, pp. 1185–230.

Hafer, R.W., Haslag, J.H. and Jones, G. "On money and output: Is money redundant?" *Journal of Monetary Economics*, 54, 2007, pp. 945–54.

Ireland, P.N. "Alternative nominal anchors." *Canadian Journal of Economics*, 31, 1998, pp. 365–84.

Judd, J.P. and Rudebusch, G. "Taylor's rule and the Fed: 1970–1997." *Federal Reserve Bank of San Francisco Economic Review*, 1998, 3, pp. 3–16.

Kahn, G.A. and Parrish, K. "Conducting monetary policy with inflation targets." *Federal Reserve Bank of Kansas City Economic Review*, 1998, pp. 5–32.

Keynes, J.M. *The General Theory of Employment, Interest and Money*. New York: Macmillan, 1936.

Levin, A., Wieland, V. and Williams, J.C. "Robustness of simple monetary policy rules under model uncertainty." In J.B. Taylor, ed., *Monetary Policy Rules*. Chicago: Chicago University Press, 1999, pp. 263–99.

Levin, A., Wieland, V. and Williams, J.C. "The performance of forecast-based monetary policy rules under model uncertainty." *Board of Governors of the Federal Reserve System*, Working Paper 2001–39, 2001.

McCallum, B.T. "On consequences and criticisms of monetary targeting." *Journal of Money, Credit, and Banking*, 17, 1985, pp. 570–97.

Nelson, E. "Direct effects of base money on aggregate demand: theory and evidence." *Journal of Monetary Economics*, 49, 2002, pp. 687–708.

Poole, W. "Optimal choice of monetary policy instruments in a simple stochastic macro model." *Quarterly Journal of Economics*, 84, 1970, pp. 197–216.

Romer, D. "Keynesian macroeconomics without the LM curve." *Journal of Economic Perspectives*, 14, 2000, pp. 149–69.

Rudebusch, G.D. and Svensson, L.E.O. "Eurosystem monetary targeting: lessons from US data." *European Economic Review*, 46, 2002, pp. 417–42.

Sims, C.A. "A review of monetary policy rules." *Journal of Economic Literature*, 39, 2001, pp. 562–6.

Taylor, J.B. "Discretion versus policy rules in practise." *Carnegie–Rochester Conference Series on Public Policy*, 39, 1993, pp. 195–214.

Taylor, J.B. *Monetary Policy Rules*. Chicago: Chicago University Press, 1999.

Walsh, C.E. *Monetary Theory and Policy*. Cambridge, MA: MIT Press, 2006.

Wang, S. and Handa, J. "Monetary policy rules under a fixed exchange rate regime: empirical evidence from China." *Applied Financial Economics*, 17, 2007, pp. 941–50.

Woodford, M. "How important is money in the conduct of monetary policy?" *NBER Working Paper* No. 13325, 2007.

14 The classical paradigm in macroeconomics

The most enduring tradition in short-run macroeconomics is represented by the classical paradigm. This chapter covers the various schools and models of this paradigm. The microeconomic foundations of the classical paradigm lie in the Walrasian model, presented in Chapter 3 above. In long-run equilibrium, the models of the classical paradigm imply full employment and full-employment output in the economy, and money neutrality, so that there is no need for the pursuit of fiscal or monetary policies to improve on the already perfect functioning of the economy. In short-run equilibrium, deviations from full employment occur due to errors in price or inflationary expectations. These deviations and the resulting departures from the neutrality of money are transient and self-correcting.

The chapter starts with the stylized facts on the relationship between money, inflation and output. To be valid and useful, macroeconomic theories need to be able to explain these facts.

Key concepts introduced in this chapter

- The neoclassical model
- The employment–output relationship/curve
- The long-run aggregate supply relationship/curve
- The short-run aggregate supply relationship/curve
- The natural rate of unemployment
- Economic liberalism
- *Laissez faire*
- The Great Depression
- Monetarism
- The traditional classical approach/theory
- The modern classical model/theory
- The new classical model
- Real business cycle theory
- Notional versus effective demand and supply functions
- Ricardian equivalence
- Pigou and real balance effects
- Rational expectations
- Expectations-augmented Phillips curve
- Friedman and Lucas supply rules

Chapter 1 presented a brief definition of the classical paradigm and its component models. Chapter 2 covered the heritage of monetary and macroeconomic theory, and in Chapter 3 we presented microeconomic analyses of the demand and supply functions for commodities, money and labor. These chapters should be reviewed at this stage.

Chapter 13, on the determination of aggregate demand, is a prerequisite for this chapter, which takes the determination of aggregate demand as a given for its analysis.

As was pointed out in Chapter 1, there are two major paradigms in short-run macro modeling: classical and Keynesian. The classical paradigm encompasses the traditional classical set of ideas, the neoclassical model, the modern classical model and the new classical model. This chapter starts with the neoclassical model and returns to the other two models towards the end.

As in earlier chapters, lower case symbols will generally refer to the real values of the variables and upper case symbols to their nominal values.

Stylized facts on money, prices and output

Monetary macroeconomics is vitally concerned with the empirical relationships between monetary policy, prices, inflation and output. For a macroeconomic theory to be valid and useful, it must do a reasonable job of explaining these relationships. The stylized facts on these relationships are:

1. Over long periods of time, there is a roughly one-to-one relationship between the money supply and inflation (McCandless and Weber, 1995).
2. Over long periods, the relationship between inflation and output growth is not significant or, if significant, is not robust (Taylor, 1996).[1] This is also so for money growth and output growth (Lucas, 1996; Kormendi and Meguire, 1984; Geweke, 1986; Wong, 2000), though some studies show a positive correlation between these variables, especially for low-inflation countries (McCandless and Weber, 1995; Bullard and Keating, 1995), while others show a negative relationship (Barro, 1996).
3. Over long periods of time, the correlation between money growth rates and nominal interest rates is very high (about 0.7 or higher) (Mishkin, 1992; Monnet and Weber, 2001), so that changes in interest rates tend to reflect changes in inflation.
4. In the short run, changes in money supply and interest rates have a strong impact on aggregate demand (Anderson and Jordan, 1968; Sims, 1972).
5. Changes in money growth lead to changes in real output in business cycles (Friedman and Schwartz, 1963a, b). Unanticipated money supply changes affect output (Barro, 1977), as do anticipated ones (Mishkin, 1982) (see Chapter 9). Negative shocks to money supply have a stronger impact on output than positive ones (Cover, 1992).
6. Monetary policy increases (decreases) in the short-term interest rate lead to a decline (an increase) in output (Eichenbaum, 1992).
7. On monetary policy dynamics in the short run, monetary shocks build to a peak impact on output and then gradually die out, so that there is a "hump-shaped pattern" of the effect of monetary policy on output (see Mosser (1992), Nelson (1998) and Christiano *et al*. (1999) for evidence on the United States, and Sims (1992) for evidence for some

1 In relation to this proposition, Taylor (1996, p. 186) concludes that "there is now little disagreement … that there is no long-run trade-off between the rate of inflation and the rate of unemployment."

other countries) with the peak effect occurring with a lag longer than one year, sometimes two to three years.

8 As a corollary of the preceding point, the impact of monetary shocks on prices occurs with a longer lag than on output, so that the impact of monetary shocks on output does not mainly occur through prior price movements.

9 For the short run, since inflation responds more gradually than output to monetary policy changes, expected inflation also responds more gradually. Therefore, errors in price or inflation expectations do not provide a satisfactory explanation of the response of output to monetary policy shifts.

10 Unanticipated price movements explain only a small part of output variability. The impact of unanticipated, as well as anticipated, monetary shifts on real output do not necessarily occur through prior price/inflation increases (Lucas, 1996, p. 679).

11 The responses of output and prices to monetary shocks differ over different episodes. They are also stronger for contractionary than for expansionary monetary episodes.

12 Contractionary monetary policies to reduce inflation do initially reduce output significantly (Ball, 1993), often for more than a year. The cost in terms of output tends to be larger if inflation is brought down gradually rather than rapidly. It is lower if the policy has greater credibility (Brayton and Tinsley, 1996).

14.1 Definitions of the short run and the long run

Equilibrium in a model is defined as that state in which all markets clear, so that the demand and supply in each market are equal. Another definition is that it is the state from which there is no inherent tendency to change. The classical paradigm uses the former definition. The latter definition is met when the former one is satisfied.

The *long run* of the short-run macroeconomic model is defined as that *analytical* period in which:

1 There are no adjustment costs, inertia, contracts or rigidities, so that all adjustments to the desired or equilibrium values of its variables are instantaneous.

2 There are no errors in expectations, so that the expected values of the variables are identical with their actual values. This condition is trivially satisfied if there is certainty.

3 There exists long-run equilibrium in all markets.

Note that these assumptions imply that there are no labor contracts between firms and workers and no price contracts among firms, so that prices and nominal and real wages adjust instantly and fully to reflect market forces. However, the long run of the short-run macroeconomic model still assumes that the capital stock, labor force and technology are constant.

Given these assumptions, the economy's resulting long-run employment level is said to be the "full-employment" level and its long-run output is said to be "full-employment output," for which our symbol is y^f. Note that the long-run – or full-employment – output is not really the maximum output that the economy could produce at any time, e.g. if all its resources were used 24 hours a day. It is also not the equilibrium level of output that would be produced in the short run or the actual output that might be produced when the economy is not in equilibrium. Hence, macroeconomics interprets the term "*fully employed*" in a special way. It is formally defined as being the level of output and employment that would exist in long-run equilibrium of the short-run macroeconomic model. Intuitively, this corresponds to the levels

of output and employment that can be sustained by the economy over the long run with its current supplies of the factors of production and its current technology – given its current economic, political and social structures, as well as the wishes of the owners of the factors of production.

By comparison with the definition of the long run, the *short run* in the context of the short-run macroeconomic model is defined as that analytical period in which:

1 Some variables, especially the capital stock, technology and labor force, are constant.
2 There can be adjustment costs, e.g. of adjusting prices, wages, employment and output to their desired levels, as well as inertia, contracts or/and rigidities of some kind.
3 There can be errors in expectations; e.g. the expected values of variables such as prices, inflation, wages and aggregate demand, differ from their actual values.
4 There can be disequilibrium in any or all of the markets.
5 The short-run equilibrium values of the various variables can differ from their long-run values. In particular, the short-run output can be greater or less than its full-employment level.

The *actual* values of the variables in the economy can differ from their long run and short-run equilibrium levels for a variety of reasons. This would occur if the actual economy is not even in short-run equilibrium or suffers short-run deviations from full employment due to factors other than errors in price expectations. Note that actual output occurs in a chronological time period whereas the short run and the long run are analytical, hypothetical, constructs. Illustrating this point by reference to the output of commodities, there are three concepts of output: actual output, short-run equilibrium output and long-run equilibrium (full-employment) output.

Also, note that the terms "short run" and "long run" are different in meaning from their counterparts of "short period" or "short term" and "long period" or "long term." The former are analytical constructs indicative of the forces allowed to work in the analysis; the latter are chronological ones and refer to an interval of actual time. In the real world, the economic forces encompassed in both the analytical short run and the long run operate simultaneously at every moment of time in the economy. To illustrate, the population and the capital stock are continuously changing, so that the analytical forces of the long-run growth models must be continuously operating in the economy, even over the next day, month or quarter.[2] At the same time, the analytical forces of the short-run macroeconomic models are also simultaneously operating in the economy.

14.2 Long-run supply side of the neoclassical model

Production function

In industrialized, as against agricultural, economies, capital and labor are the dominant inputs in production, while land plays only a minor role and is normally not included in macroeconomic analysis. With this assumption, the production of commodities is taken to

2 Note that Keynes's famous remark that "in the long run we are all dead" is not valid under the above interpretation of the long run.

use only labor and capital as inputs. The production function for the economy – as represented by that for the representative firm – can then be written as:

$$y = y(n, K) \quad y_n, y_K > 0; y_{nn}, y_{KK} < 0 \tag{1}$$

where:

y = output
K = physical capital stock
n = labor employed.

The physical stock of capital has already been assumed to be constant in the short-run macroeconomic context of our analysis, so that $K = \underline{K}$, where the underlining indicates "constancy" or "exogeneity," as required by the definition of the short run. Hence, we have:

$$y = y(n, \underline{K}) \quad y_n > 0, y_{nn} < 0 \tag{2}$$

With this modification, labor is left as the only variable input, with a positive relationship between output and employment. The assumption of $y_n > 0$ and $y_{nn} < 0$ is that the marginal productivity of labor is positive but diminishing: successive increments of labor yield smaller and smaller increments of output.

Labor market in the long run

The specification of the labor market requires specification of the demand and supply functions of labor and its equilibrium condition. Their derivation was presented in Chapter 3. We here present the simplified derivations used in standard neoclassical macroeconomic models, which imply that the demand and supply of labor depend only on the real wage rate. However, intertemporal analysis implies that both these functions will also depend on the real interest rate and future wage rates. On this issue, empirical studies of both labor demand and supply analysis show that, for short periods, neither significantly depends on the interest rate or the future wage rates for the ranges in which these variables normally fluctuate. Therefore, allowing empirical findings to determine the relevant theoretical assumptions, the following macroeconomic model has labor demand and supply functions that depend only on the (current) real wage.

Production analysis assumes that firms maximize profits and operate in perfectly competitive markets. Hence, they employ labor until its marginal revenue product equals its nominal wage rate. Using the price level to divide both of these variables, in perfect competition and for the representative firm, profit maximization requires that firms employ labor up to the point where the real value of its marginal product equals its real wage rate. That is,

$$y_n(n, \underline{K}) = w \tag{3}$$

where:

y_n = marginal product of labor
w = real wage rate.

Since K is held constant at $\underline{K}$, it is omitted from further analysis. Solving (3) for employment n, and designating this value as the demand for labor n^{d} by firms:

$$n^{\mathrm{d}} = n^{\mathrm{d}}(w) \quad \partial n^{\mathrm{d}}/\partial w < 0 \tag{4}$$

Chapter 3 derived the supply function of labor from utility maximization subject to a budget constraint. Its simple version in a single commodity world is:

$$n^s = n^s(w) \quad \partial n^s/\partial w > 0 \tag{5}$$

where n^s is the supply of labor. Note that (5) specifies that the supply of labor depends upon the real rather than the nominal wage. Hence, workers are free from *price illusion*, which is the distortion caused when money wage rates and the prices of commodities rise by identical proportions but the workers, looking at the rising nominal wage rates, believe that they are better off even though the purchasing power of wages has remained unchanged.

Long-run equilibrium levels of employment and output

It is important to note that there are no adjustment costs or errors in expectations in the preceding framework, so that its equilibrium will be the long-run equilibrium, as against a short-run equilibrium in a model that allows for adjustment costs and expectational errors.

Equilibrium in the labor market requires that:

$$n^d(w) = n^s(w) \tag{6}$$

Since (6) is an equation in only one variable, w, solving it would yield the long-run equilibrium wage rate w^{LR}. This wage rate, substituted in either the demand or the supply function, yields the long-run equilibrium level of employment n^{LR}. This level of employment, substituted in the production function (2), yields the long-run equilibrium level of output y^{LR} for the economy.

Note that n^{LR} equals both the demand and supply of labor at the equilibrium wage. Therefore, at n^{LR}, all the workers who want jobs at the existing wage rate are employed and the firms can get all the workers that they want to employ. This is the definition of full employment,[3] so that the equilibrium level of employment n^{LR} represents full employment and, to emphasize this property, can be designated as n^f. Its corresponding long-run equilibrium output level y^{LR} is the full-employment level of output y^f.

These conclusions on the labor market equilibrium are that:

$$n = n^{LR} = n^f \tag{7}$$

From (2) and (7), we have:

$$y = y^{LR} = y^f \tag{8}$$

Diagrammatic analysis of output and employment in the neoclassical model

Figure 14.1 plots the demand and supply functions of labor, with the usual slopes for demand and supply curves. Equilibrium occurs at (n^{LR}, w^{LR}). Figure 14.2 plots the

3 The definition of full employment is: it is the level of employment such that all the workers who want jobs at the going wage rate have a job, excluding those who are temporarily between jobs during the job search process.

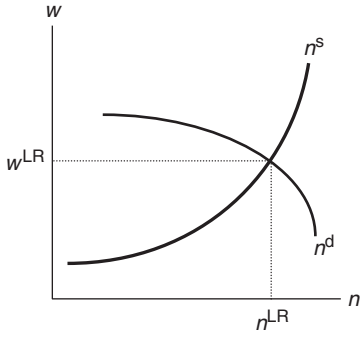


Figure 14.1

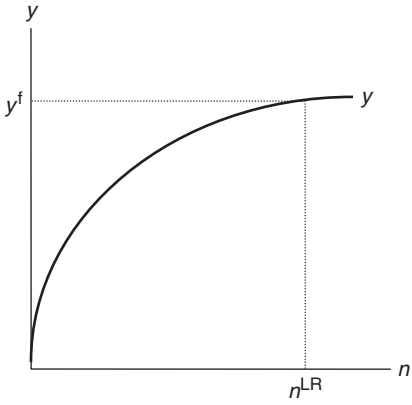


Figure 14.2

production function: output y is plotted against employment n and the curve is labeled y (output). This curve has a positive slope and is concave, representing the assumption of the diminishing marginal productivity of labor. The equilibrium level of employment n^{LR} determined in Figure 14.1 is carried over into Figure 14.2 and implies the equilibrium – and profit-maximizing – output y^f.

From (7) to (8) – and as shown in Figures 14.1 and 14.2 – the equilibrium levels of the real wage rate, employment and output do not depend upon the price level. They are determined uniquely and, in particular, are independent of the demand for commodities. The factors that will change these equilibrium values w^{LR}, n^{LR} and y^f are shifts in the production function – with implied shifts in the demand for labor – and in the supply of labor. Other shocks to the economy which do not bring about these shifts will not alter w^{LR}, n^{LR} and y^f. Among these shocks are changes in the monetary and fiscal policy variables, which change aggregate demand but do not appear as arguments of the production function and the supply function of labor. This is a very strong implication for the equilibrium states of the neoclassical model. It implies that aggregate demand policies, such as monetary and fiscal policies, cannot affect the equilibrium levels of wages, employment and output in the economy. This implication will be discussed below at greater length.

The irrelevance of aggregate demand for equilibrium output and employment and its empirical validity

Since y^f is independent of the demand side of the economy, the preceding analysis implies that:

1 Aggregate demand and shifts in it cannot change the equilibrium levels of output, employment, real wage and other real variables. Aggregate demand is irrelevant to their determination.
2 Since the model relies on equilibrium for its results, it has the implicit assumption that the economy has adequate and sufficiently fast-reacting equilibrating mechanisms to force aggregate demand y^d into equality with y^f continuously or in a short enough period for the deficit or excess of demand not to affect the production and employment decisions of firms and/or the consumption demand and labor supply decisions of households. Therefore, we might caricature the adjustment process as: "the equilibrium level of supply creates its own demand"[4] through the equilibrating variations in prices, wages and interest rates. This is quite a strong assumption and not all economies in all possible stages of development or of the business cycle meet it.[5]

The irrelevance of aggregate demand and, by implication, of monetary and fiscal policies, which can change that demand, for the determination of output and unemployment is an extremely strong implication of the equilibrium properties of the neoclassical model. Comparison of this implication with the stylized facts given at the beginning of this chapter shows that the implication is clearly not valid. Therefore, in the preceding part of the neoclassical model, either the equilibrium assumption must be abandoned, or its specifications of the production process or of labor demand and labor supply have to be modified. As discussed later, the Lucas supply analysis makes this modification for the production process, while Friedman's expectations-augmented analysis does so for labor demand and supply.

14.3 General equilibrium: aggregate demand and supply analysis

Chapter 13 and this chapter have so far specified the markets for commodities, money and labor. The foreign exchange market has been taken to be in equilibrium through appropriate changes in the exchange rate. We have not specified the market for bonds, which are defined as non-monetary financial assets, even though this is one of the four goods in the macroeconomic model. This omission is justified by Walras's law, which specifies that in a four-good economy, if three of the goods markets are in equilibrium the market for the fourth good must also be in equilibrium. Therefore, at the general equilibrium (y^{LR}, r^{LR}, P^{LR}) in the preceding analysis, the bond market will also be in equilibrium and can be omitted from explicit consideration.

A full view of the economy requires simultaneous consideration of all markets, and general equilibrium in the economy implies a simultaneous solution to the equilibrium equations for all the three sectors. We consider this in two alternate ways, demand–supply analysis and the IS–LM analysis.

4 This sounds similar to but is really weaker than "Say's law," which is discussed in Chapter 17.
5 If this assumption is not valid for an economy at any stage, then the equilibrium properties of the neoclassical model will not be applicable. The pertinent properties will be those of its disequilibrium analysis.

Demand–supply analysis

The aggregate supply equation derived so far is:

Aggregate supply:

$$y^{s} = y^{LR} = y^{f} \tag{9}$$

The aggregate demand equation is the one derived from either the IS–LM analysis or the IS–IRT analysis of the preceding chapter. Since there are two alternate aggregate demand equations, we use only their general form, stated as:

$$y^{d} = y^{d}(P; g, \theta) \tag{10}$$

where g is the vector of fiscal policy variables and θ is the relevant monetary policy variable. Note that (10) assumes that Ricardian equivalence does not hold. If it does so, then, as shown in the preceding chapter, fiscal variables will not be in the aggregate demand function, so that the AD function will become $y^{d}(P; \theta)$. The validity of Ricardian equivalence is doubtful, so we proceed with the aggregate demand function $y^{d}(P; g, \theta)$.

Equilibrium in the commodity market requires that:

$$y^{s} = y^{d}(P; g, \theta) \tag{11}$$

(9) and (11) have two endogenous variables: y and P. Of these two equations, (9) clearly determines y, even without reference to (10), as being equal to y^{LR}. Therefore, the aggregate demand equation (10) can only determine P, with y on its left side being set equal to y^{LR}.

Aggregate demand and supply curves

The above conclusion is illustrated in Figure 14.3. Equation (9) implies that the *long-run* aggregate supply curve LAS is vertical while, as shown in the preceding chapter, (10) implies that the open economy aggregate demand curve AD has a negative slope. An examination of this figure clearly shows that shifts in the aggregate demand curve will not change the equilibrium output but only the price level, while changes in the

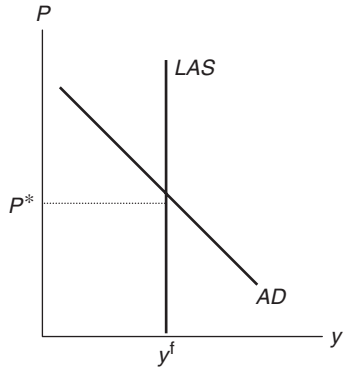


Figure 14.3

aggregate supply will change both output and price level. This is a very strong conclusion and is, as we shall see later in this and the next three chapters, at the heart of the debate on the ineffectiveness of monetary and fiscal policies for the equilibrium states of the neoclassical model and its related modern classical and new classical models.

Equilibrium output and prices

Equations (9) to (11) determine the long-run equilibrium values of output y and price level P from:

$$y^{\mathrm{LR}} = y^{\mathrm{f}} \tag{12}$$

$$P = f(g, \theta; y^{\mathrm{LR}}) \tag{13}$$

so that monetary and fiscal policies can change aggregate demand and the price level, but not long-run output. In the presence of Ricardian equivalence, $P = f(\theta, y^{\mathrm{LR}})$, so that changes in monetary policy and in y^{LR} can change the price level, but fiscal policy cannot do so.

In terms of the general equilibrium neoclassical model, (12) determines output and (13) determines the price level. These equations imply the aggregate demand, monetary and fiscal multipliers on output as:

$$\partial y^{\mathrm{LR}}/\partial y^{\mathrm{d}} = 0, \quad \partial P^{\mathrm{LR}}/\partial y^{\mathrm{d}} > 0 \tag{14}$$

$$\partial y^{\mathrm{LR}}/\partial g = 0, \quad \partial P^{\mathrm{LR}}/\partial g \geq 0 \tag{15}$$[6]

$$\partial y^{\mathrm{LR}}/\partial \theta = 0, \quad \partial P^{\mathrm{LR}}/\partial \theta > 0 \tag{16}$$

which clearly indicate that the long-run equilibrium level of output is not responsive to monetary policies, whereas that of the price level is responsive to these policies.[7] This also applies to fiscal policies, except under Ricardian equivalence when fiscal policy affects neither aggregate demand nor output nor the price level.

Supply shifts

The impact of a change in output on the price level, represented by $\partial P/\partial y^{\mathrm{LR}}$, is negative in both the IS–LM and IS–IRT models, though its magnitude depends on whether the money supply or the interest rate is the exogenous monetary policy variable. The basic reason for the negative impact of output increases on the price level is that an increase in output raises the transactions demand for money, which has to be offset by lower prices. We leave it to the interested reader to derive $\partial P/\partial y$ for the IS–LM and IS–IRT models, using the information given in this and the last chapter.

6 $\partial P^{\mathrm{LR}}/\partial g = 0$ if Ricardian equivalence holds.

7 This model can also be used to derive the quantity theory proposition that, in equilibrium, a change in the money supply causes a proportionate change in the price level, even though the neoclassical model encompasses the interest sensitivity of money demand – and therefore allows for a positive speculative demand for money.

14.4 Iterative structure of the neoclassical model

Another procedure for studying simultaneous equilibrium in all sectors of the economy emerges if the final *equilibrium* equation for each of the sectors is examined separately in the overall problem. The following equations (17) to (22) incorporate information on all the sectors.

Production–employment sector:

$$y = y^{\mathrm{f}} \tag{17}$$

The commodity market IS equation:

$$y = y(r, P; g, \theta) = \left(\frac{1}{1 - c_y + c_y t_y + \dfrac{1}{\rho^{\mathrm{r}}} z_{cy}(1 - t_y)} \right) \cdot \left(\left\{ c_0 - c_y t_0 + i_0 - i_r r + g + x_{c0} - x_{c\rho} \rho^{\mathrm{r}} \right\} + \frac{1}{\rho^{\mathrm{r}}} \left\{ -z_{c0} + z_{cy} t_0 - z_{c\rho} \rho^{\mathrm{r}} \right\} \right) \tag{18}$$

Fisher equation:

$$R = (1 + r)(1 + \pi^{\mathrm{e}}) \tag{19}$$

From the employment–output sector equilibrium, we know that the long-run equilibrium output is y^{f}. Substitute this equilibrium level of output into the IS relationship (18), which yields the long-run equilibrium real rate of interest as $r^{\mathrm{LR}}{}_0$:

$$r^{\mathrm{LR}}{}_0 = \frac{1}{i_r} \left(\left(c_0 - c_y t_0 + i_0 + g_0 + x_{c0} - x_{c\rho} \rho^{\mathrm{r}} \right) + \frac{1}{\rho^{\mathrm{r}}} \left(-z_{c0} + z_{cy} t_0 - z_{c\rho} \rho^{\mathrm{r}} \right) \right) - \frac{\left(1 - c_y + c_y t_y + \dfrac{1}{\rho^{\mathrm{r}}} z_{\mathrm{cy}} (1 - t_y) \right)}{i_{\mathrm{r}}} y^{\mathrm{f}} \tag{20}$$

We now know the long-run equilibrium level of the real interest rate, even without introducing the monetary sector into the analysis so far.

Money market equilibrium depends on whether the central bank uses the interest rate or the money supply as the exogenous monetary policy variable. The relevant equations are:

LM equation in the IS–LM model:

$$\overline{M}/P = m_y y + (FW_0 - m_R R) \tag{21}$$

Money market equilibrium condition in the IS–IRT model:

$$M/P = m_y y + (FW_0 - m_R \overline{R}) \tag{22}$$

The following analysis deals separately with each of these cases.

Determination of the price level for an exogenously given money supply (the IS–LM model)

The preceding results imply that the quantity of the money supply is irrelevant to the determination of the long-run equilibrium values of both output and the real rate of interest. To determine the price level in the neoclassical model based on the IS–LM analysis, we start with the Fisher equation, which determines the nominal interest rate from $(1+r^{\mathrm{LR}}{}_0)(1+\pi^{\mathrm{e}})$, so that, depending on the expected inflation rate, different nominal interest rates are consistent with $r^{\mathrm{LR}}{}_0$.

In the LM equation with an exogenously given money supply, substituting the long-run equilibrium level of income y^{f} and the nominal interest rate from the Fisher equation as approximately equal to $(r^{\mathrm{LR}}{}_0+\pi^{\mathrm{e}})$ yields the equilibrium price level as:

$$P=\frac{\alpha\cdot i_r\cdot M_0}{m_R\left[y^{\mathrm{f}}-\alpha\cdot\left(\left\{c_0-c_y t_0+i_0-\dfrac{i_r}{m_R}FW_0+i_r\pi^{\mathrm{e}}+g_0+x_{c0}-x_{c\rho}\rho^{\mathrm{r}}\right\}+\dfrac{1}{\rho^{\mathrm{r}}}\left\{-z_{c0}+z_{cy}t_0-z_{c\rho}\rho^{\mathrm{r}}\right\}\right)\right]} \tag{23}$$

where:

$$\alpha=\left(\frac{1}{1-c_y+c_y t_y+\dfrac{1}{\rho^{\mathrm{r}}}z_{cy}(1-t_y)+i_r\dfrac{m_y}{m_R}}\right)$$

Note the sequence in the above procedure. The production–employment sector alone determines the full-employment output in (17), without any reference to the interest rate and the price level; the expenditure sector then determines the long-run equilibrium real interest rate, without any reference to the price level or the monetary sector; but the price level is determined by reference to all the sectors of the economy.

In the special quantity theory case when $m_R=0$, the monetary sector condition with $y=y^{\mathrm{f}}$ simplifies to:

$$M/P=m_y y^{\mathrm{f}}$$

which can be rearranged as:

$$P=\left(\frac{1}{m_y y^{\mathrm{f}}}\right)M \tag{24}$$

which does not include either r or R as a variable, so that we need to know only the money supply and output to determine the price level. This can be taken to be a modern derivation of the quantity theory, as in Pigou's version in Chapter 2. In it, the dependence of the price level upon the interest rate is eliminated by the assumption that the interest sensitivity of money demand is zero and the economy has full-employment output.

Determination of the price level for an exogenously given interest rate (the IS–IRT model)

In the IS–IRT model of the preceding chapter, it was assumed that the central bank sets the real interest rate, with its level specified as $r^T{}_0$. Aggregate demand y^{d} is then given by the commodity market at the given real interest rate. This AD equation is:

$$y^{\mathrm{d}} = y(r_0^{\mathrm{T}}, P; g, \theta) = \left(\frac{1}{1 - c_y + c_y t_y + \frac{1}{\rho^{\mathrm{r}}} z_{cy}(1 - t_y)} \right) \cdot \left(\left\{ c_0 - c_y t_0 + i_0 - i_r r_0^{\mathrm{T}} + g + x_{c0} - x_{c\rho} \rho^{\mathrm{r}} \right\} + \frac{1}{\rho^{\mathrm{r}}} \left\{ -z_{c0} + z_{cy} t_0 - z_{c\rho} \rho^{\mathrm{r}} \right\} \right) \tag{25}$$

Given output supply at y^{f}, and the real interest rate at r^{T}, this equation determines the price level P.

Policy implications: the ineffectiveness of monetary and fiscal policies in changing output and employment

Important policy implications follow from the iterative nature in the long-run equilibrium of the neoclassical model. No matter what is the monetary policy variable, neither monetary nor fiscal policies can affect equilibrium output, since neither appears in (17). Therefore, these policies are useless for increasing the long-run equilibrium levels of employment and reducing the equilibrium level of unemployment. However, with the economy in long-run equilibrium, these policies are not only useless, they are also not needed since the long-run equilibrium employment is at full employment. Hence, in the long run, there is *no need or scope* for the authorities to pursue policies to increase employment or output, which are both at their full-employment levels. Any such attempt will be ineffective.

If we compare these implications on the irrelevance of monetary and fiscal policies for output and employment with the stylized facts, these implications are clearly invalid. Expansionary monetary and fiscal policies do increase aggregate output and lower unemployment, and the reverse is true for contractionary policies. Therefore, the preceding model needs to be modified in order to be valid and useful. The classical economists do this, as explained later, by introducing uncertainty, with errors or misperceptions in expectations, into the model.

14.4.1 The rate of unemployment and the natural rate of unemployment

The level of unemployment is defined as:

$$U = L - n \tag{26}$$

where:

U = level of unemployment
L = labor force.

Since $n \leq L$, unemployment is always non-negative. Assuming L to be exogenously given as $\underline{L}$, so that it does not vary with the real wage, the labor force will be the sum total of workers who are able and willing to work at *any* wage. In this case, L represents the maximum amount of potential employment in the economy. But if $L = L(w)$, it is likely that the number of workers willing to work increases as real wages rise, so that $L' \geq 0$. For our fairly basic analysis at this point, we will make the former assumption.

The natural rate of unemployment

The *long-run equilibrium level* of unemployment U^{LR} is:

$$U^{\mathrm{LR}} = L - n^{\mathrm{LR}} \tag{27}$$

The *long-run equilibrium rate* of unemployment u^{n}, with the superscript n standing for "natural," which itself stands for long-run equilibrium, is:

$$u^{\mathrm{n}} = U^{\mathrm{LR}}/L = 1 - n^{\mathrm{LR}}/L \tag{28}$$

From (17), since aggregate demand and its determinants cannot change output, they also cannot change unemployment. Hence, from (28),

$$\partial u^{\mathrm{n}}/\partial y^{\mathrm{d}} = \partial u^{\mathrm{n}}/\partial \theta = \partial u^{\mathrm{n}}/\partial g = 0$$

where θ is the monetary policy variable and g is the fiscal variable. Hence, in the neoclassical model, since n^{LR} is independent of the demand side of the economy and L is exogenous, u^{n} is also independent of changes in aggregate demand and, therefore, of monetary and fiscal policies. Note that $u^{\mathrm{n}} > 0$ by virtue of structural, frictional, search and seasonal unemployment in the economy, which prevent the employment of all members of the labor force since some would have inappropriate skills and education, be in inappropriate locations or require wages in excess of their marginal productivity in the current state of the economy.

On the characteristics of the natural rate of unemployment in the neoclassical model, this rate cannot be changed by monetary or fiscal policies (Friedman, 1977). It does, however, depend upon the supply structure – the labor market relationships and the production function – of the economy and will change as the supply structure changes. Technical change and changes in educational and skill requirements, the level of education of the labor force, the availability of information on jobs and workers, the location of industry, etc., are thus likely to change the natural rate of unemployment. This rate is therefore itself a variable, though not one that can be changed by demand shifts in the economy, including the pursuit of monetary and fiscal policies. Shifts in the supply side of the economy can change the natural rate. Among these shifts is technical change and shifts in the industrial structure due, among other things, to shifts in the structure of demand among the sectors of the economy.

The natural rate of unemployment rises during a transition from one industrial–agricultural structure of the economy to a different one. Suppose that industry A is declining and laying off workers while industry B is expanding its labor force. The process of transfer of workers involves searching for new jobs by the laid-off workers, so that search unemployment increases during the transition. Further, some of the laid-off workers may possess skills not needed in industry B and may become permanently unemployed. This increases structural

unemployment in the economy. That is, the shift in the economy's industrial structure induces a *transitional* increase in the natural rate of unemployment, but it can also imply a long-run shift in that rate.

14.4.2 IS–LM version of the neoclassical model in a diagrammatic form

Figure 14.3 brings together the information in equations (17) to (22). (18) to (22) specify the downward-sloping AD curve. From (17), since output does not depend on P, the aggregate supply (AS) curve is vertical at y^f. It is often called the "full-employment (y^f) curve." Equilibrium in all the sectors of the economy exists at the point (P^*, y^f).

An expansionary monetary policy with an exogenous money supply (the IS–LM model)

Figure 14.4 illustrates the IS–LM model in the (r, y) space. Equation (17) specifies the LAS curve at the full-employment output y^f, (18) specifies the IS curve, and (19) and (21) are used to specify the LM curve. The intersection of the IS and LM curves determines the level of aggregate demand. General equilibrium in the commodity, money and output sectors requires that all three curves intersect at the same point. This is the case shown in Figure 14.4.

We next provide some examples of how the IS–LM analysis and diagram can be manipulated for comparative static studies. This is done in Figure 14.5 for monetary policy. Suppose that the economy is initially in overall equilibrium at the point a and the money supply increases, shifting the LM curve from LM_0 to LM_1. The new equilibrium between the monetary and expenditure sectors is shown by the point d and represents nominal aggregate demand. But output specified by the output–employment sector is at y^f. Since aggregate demand at d exceeds the supply of output y^f, prices rise. As P rises, the LM curve shifts towards the left. However, the rise in P leaves the IS and AS curves unchanged. Prices will then continue to rise as long as aggregate demand for output exceeds its supply. This will occur until the leftward shifts in the LM curve due to the price increases take it sufficiently back (i.e. from LM_1 to LM_1') to pass through the initial equilibrium point a. In short, in terms of equilibrium states, an increase in the money supply increases

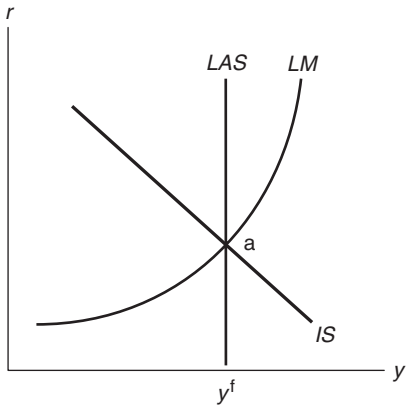


Figure 14.4

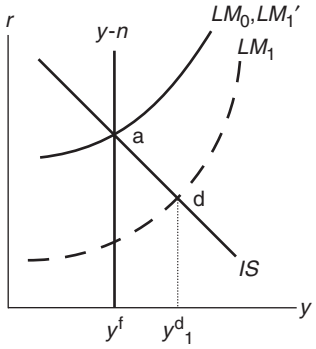


Figure 14.5

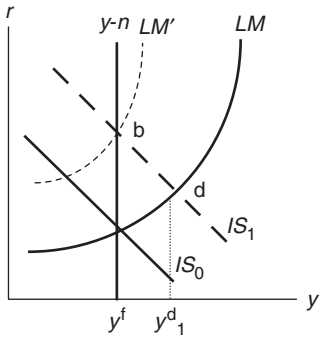


Figure 14.6

aggregate demand, without affecting output, and raises prices to a new level, with the increase in the price level being proportionate to the increase in the money supply. The interest rate is unchanged between the old and the new equilibrium positions. Comparing these implications with the stylized facts given at the beginning of this chapter shows that they are not valid for the short run. Therefore, the model presented so far needs to be modified.

An expansionary fiscal policy with an exogenous money supply

Now suppose that fiscal policy had become expansionary with an increase in government expenditures. This would have increased the real aggregate demand for commodities and shifted the IS curve from IS_0 to IS_1 in Figure 14.6. Nominal aggregate demand at the point d (at the intersection of the IS_1 and LM curves) exceeds output y^f. Prices rise, eventually lifting the LM curve to a new overall equilibrium at point b, with the final LM curve as LM'. Both prices and interest rates are higher in the new equilibrium position at b than at the initial one at point a. Hence, increases in fiscal expenditures increase the equilibrium levels of interest rates and prices. However, they do not change output. The last implication is clearly not valid for the short run.

14.5 Fundamental assumptions of the Walrasian equilibrium analysis

The preceding analysis has focused on the long-run equilibrium states of the model and is a succinct macroeconomic version of the *Walrasian model in the absence of uncertainty*. This equilibrium analysis makes four fundamental assumptions, plus a fifth one that specifies that the findings are for the certainty case. These are:

(I) *Flexible prices and wages, and the stability of markets*
The prices of all the goods in the economy are assumed to be flexible and adjust to equate demand and supply in the relevant market. They increase if there is excess demand and decrease if there is excess supply. These prices include wages, which is the price of labor. Wages are flexible in both nominal and real terms.

(II) *Perfect market hypothesis*[8]
Each market has perfect competition and clears *continuously*, so that we can focus on the study of competitive general equilibrium in the economy and its properties, while largely ignoring the disequilibrium values of the variables.

(III) *Transparency of equilibrium prices*
All agents, in making their demand and supply decisions, assume that such market clearance will occur instantly after any disturbance and know or anticipate (or are informed by an agency such as a "Walrasian auctioneer" or "market coordinator") the prices at which it will occur. Further, all agents plan to produce, consume, demand money and supply labor at *only* these equilibrium prices.[9]

(IV) *Notional demand and supply functions*
All economic agents assume that they will be able to buy or sell as much as they want to at the long-run equilibrium prices. The demand and supply functions derived under this assumption are known as *notional* (as against effective) *demand and supply functions*.

The Walrasian-based models now also include:

(V) *Assumptions on uncertainty*
If the model assumes certainty, its results provide the long-run equilibrium of the economy. Its extension to include uncertainty will produce deviations from this long-run equilibrium, with the nature of the deviations depending on the way uncertainty is handled in the extended model. When the equilibrium of the economy under uncertainty differs from the long-run equilibrium of the model, such an equilibrium is designated as a short-run equilibrium. One of the causes of such a deviation from the long-run equilibrium is errors in price expectations. However, there can also be many other causes of such deviations.

The current versions of the Walrasian-based macroeconomic models assume the rational expectations hypothesis, with the long-run general equilibrium outcomes used to specify the rationally expected values of the relevant variables.

The classical equilibrium analysis of the economy and its policy recommendations require the applicability of the above fundamental assumptions: flexible prices and wages, continuous

8 A market is perfect if it has perfect competition and is fully efficient, which is defined as continuous/instantaneous market clearance. Therefore, a market can be competitive but have slow adjustment of prices, so that it is not efficient.

9 In particular, if there are delays in establishing these prices or communicating their knowledge, it is assumed that all production and trade are put on hold during the delay and that any such delay does not impose costs on economic agents. This is known as recontracting.

market clearance, transparency of equilibrium prices and notional demand and supply functions. Each assumption is related to the others but is still distinct. Any one or more of these assumptions may not be relevant to or valid for a particular stage or at a particular time in an economy.

As shown above, monetary policy is neutral in the long-run equilibrium states of the neoclassical model since changes in monetary policy affect only the nominal but not the real variables. Hence, the neoclassical model implies that, in its long-run equilibrium, monetary policy cannot change output and employment in the economy. In fact, with the economy at full employment, there is also no need for the pursuit of monetary policy. Conversely, if the above assumptions hold, changes in the money supply also cannot be detrimental for long-run output and employment in the economy. In particular, decreases in the money supply will not decrease output and employment and force the economy into a recession. Monetary policy is benign (harmless) in such a context. Note that these are assertions about the long-run analytical state, not about the short-run equilibrium or disequilibrium.

14.6 Disequilibrium in the neoclassical model and the non-neutrality of money

Note that the neoclassical model does not assert that its long-run equilibrium must always exist – as if it were an identity – and therefore it allows the possibility that the economy can be sometimes in short-run equilibrium or in disequilibrium. Further, for the study of disequilibrium to be a potentially useful exercise requires the belief that the economy will be away from its full-employment equilibrium state for significant periods of time. Intuitively, continuous general equilibrium requires, for example, the belief that an increase in the money supply *immediately* causes a proportionate increase in the price level and that a decrease in the money supply does not cause a recession in output and employment. There is considerable evidence to show that these requirements are not always, or most of the time, met in most real-world economies. Nor did the major proponents in the traditional classical and neoclassical tradition claim that they did. Among those who allowed for the existence of disequilibrium for significant periods of time, and the impact of money supply changes on output (i.e. the non-neutrality of money) during disequilibrium, were Hume, Marshall, Fisher and Pigou in the traditional classical school, and Friedman and the St Louis monetarists in the neoclassical one.

However, when the classical economists allowed for disequilibrium, they maintained that any such state of disequilibrium incorporates certain forces brought about by markets (not firms) that will force the economy into equilibrium. Among these forces are price changes and the Pigou effect, including the real balance effect. These were touched upon in Chapter 3 and will be explored more fully in Chapter 18. We explain them again, though briefly, in the following subsection as a reminder and for completeness of the neoclassical model.

14.6.1 Pigou and real balance effects

The *Pigou effect* is associated with the contributions of A.C. Pigou (at Cambridge University in England in the first half of the twentieth century) in the debate between the traditional classical school and Keynes in the 1930s. The Pigou effect is another name for the effect of

real wealth (defined as including all financial assets) on (real) consumption resulting from a change in the price level. That is, the Pigou effect is represented by:

$$\text{Pigou effect} = [\partial c/\partial \underset{+}{\text{real wealth}}] \cdot [\partial \underset{-}{\text{real wealth}}/\partial P] < 0$$

The Pigou effect works in the following manner. A disequilibrium with deficient demand for commodities will cause a fall in their prices. Since the household's wealth includes financial assets, this fall in the price level will increase the household's wealth, which, in turn, will cause consumption to rise. The latter will bring about an increase in aggregate demand in the economy. This process will continue until the demand deficiency is eliminated – that is, until the economy returns to equilibrium.[10]

The *real balance effect* is associated with the contributions of Don Patinkin during the 1940s to the 1960s and represented a refinement of the neoclassical model for the analysis of disequilibrium (Patinkin, 1965). This effect represents the impact of changes in the price level on consumption through changes in the real value of money holdings. It works as follows. A price fall due to a demand deficiency will increase the real value of money holdings and thereby increase the household's wealth. This will lead to an increase in consumption and therefore in aggregate demand. The real balance effect will continue until the demand deficiency and its associated price level decreases are eliminated.

That is, the real balance effect is represented by:

$$\text{Real balance effect} = [\partial c/\partial \underset{+}{(M/P)}] \cdot [\partial (M/P))/\partial P] < 0$$

Hence, the real balance effect and the Pigou effect are equilibrating mechanisms of the neoclassical model and require flexible prices. They were shown to imply that a fall in aggregate demand would produce a fall in prices, which would increase aggregate demand because of an increase in the real value of financial assets in the case of the Pigou effect and of money balances in the case of the Patinkin effect. Their *analytical* relevance, though under the *ceteris paribus* clause, is significant and beyond dispute. However,

> [Pigou himself] described what later came to be called the "The Pigou Effect" as a mere toy, based on "so extremely improbable assumptions" as to never be played "on the checkerboard of real life". … It would not work in any except the most formal version of a most naïve model [because, outside its *ceteris paribus* assumptions, a fall in aggregate demand would also bring about]… simultaneous bankruptcies and deflation [which] keep shifting both the LM and IS functions, and therefore the aggregate demand function towards the origin. The result is much more likely to be a depression rather than full employment.
>
> (Pesek, 1988, pp. 6–7).

10 Pigou himself was not enamored of the empirical relevance of the analytical Pigou effect as specified above. He considered that the deflationary fall in prices would be accompanied by bankruptcies among firms and a fall in the physical capital stock in production. This would cause a rise in unemployment and a fall in aggregate demand.

Further, the analyses of the real balance and the Pigou effects do not, a priori, provide any guidance on the time the neoclassical economy will need to return to long-run equilibrium under their impetus. In particular, the real balance effect can be quite weak, so that the neoclassical economy could react to an exogenous fall in demand by a very slow movement towards long-run equilibrium and, therefore, remain away from it for a significant period. Hence, it is important to analyze the short run and disequilibrium properties of the neoclassical model and derive its policy implications.

Macroeconomic theory uses two main channels for the effects of money supply changes on aggregate demand. One of these is because changes in the money supply change wealth and real balances, which changes consumption expenditures (the direct transmission channel). The other is through the effect of money supply changes on interest rates, which changes investment expenditures (the indirect transmission channel). The direct effects are considered to be relatively small over the business cycle, so that most macroeconomic models, including the popular IS–LM one, ignore them altogether.[11] Consequently, these models embody only the indirect transmission channel of monetary policy effects through interest rates.

14.6.2 Causes of deviations from long-run equilibrium

The actual economy may not be in or close to its long-run equilibrium because of:

1 Errors in expectations, either in the commodity or/and in labor markets. The analysis of this scenario is presented in the next section on the short-run equilibrium of the neoclassical model.
2 Costs of adjusting prices, wages, employment and output. The analyses of these various scenarios are presented in the next chapter on the Keynesian paradigm.
3 The absence of a mechanism for instantly restoring equilibrium. Note that the assumption of perfect competition does not a priori specify the chronological time needed by the "invisible hand of competition" to return an economy, following a shock, to its full-employment equilibrium.[12] Along with the absence of a mechanism for instantly restoring the full-employment equilibrium, the competitive economy also does not have an instantaneous mechanism operating in disequilibrium for computing the new price level and informing all firms and households of the new prices of the products. Further, there is no guarantee given to the firms that they could sell all they wanted at these prices, nor is there any guarantee that the workers will receive or recoup the incomes lost while they were unemployed. Combining this lack of a guarantee with the plausible possibility that firms and households may respond faster than markets to disequilibrium and will do so on the basis of expectations of the quantity demanded and jobs available, could produce a disequilibrium path that keeps the economy out of equilibrium for quite some (chronological) time. This scenario is sketched in the next chapter.
4 Firms are in monopolistic competition with sticky prices (see Chapter 15).

11 This implies that consumption is independent of money balances, which requires that the utility function over goods be separable between commodities and money balances. Empirically, Ireland (2001a) finds little support for a non-separable utility function for US data.

12 Because of this, the Walrasian and neoclassical models were buttressed with some *deus ex machina* such as the Walrasian auctioneer, tâtonnement (i.e. the process of "groping" towards equilibrium) and recontracting, which solved the problem analytically but represented an escape from answering the relevant question of how long it takes the real-world economies to return to equilibrium.

14.7 The relationship between the money supply and the price level: the heritage of ideas

The basic comparative static conclusions of the neoclassical macroeconomic model of this chapter were presented several centuries ago. The following quote illustrates them from the writings of David Hume, one of the founders of classical economics.

> Money is nothing but the representation of labor and commodities, and serves only as a method of rating or estimating them. Where coin is in greater plenty – as a greater quantity of it is required to represent the same quantity of goods – it can have no effect, either good or bad, taking a nation within itself; any more than it would make an alteration in a merchant's books, if instead of the Arabian method of notation, which requires few characters, he should make use of the Roman, which requires a great many.
>
> (Hume, *Of Money*, 1752).

This quote is an assertion of the basic proposition of the quantity theory of money: an increase in the money supply causes a proportional increase in prices. Further, Hume asserts that changes in the supply of money do not change real output in the economy but correspond to a change in the unit of account. The quantity theory was presented in Chapter 2.

The conclusions of this theory on the proportionate relationship between the money stock and prices, and the inability of the monetary and fiscal authorities to control output and employment, apply in equilibrium. They do not necessarily apply in disequilibrium – that is, during the adjustment from one equilibrium to another. In fact, many supporters of this theory, in Hume's time and down to the present, have viewed changes in the money stock as exerting a powerful influence on output, employment and other variables in the adjustment process. Hume himself described this process as follows.

> Notwithstanding this conclusion, which must be allowed just, it is certain that since the discovery of the mines in America, industry has increased in all the nations of Europe, except in the possessors of those mines; and this may justly be ascribed, among other reasons, to the increase of gold and silver. Accordingly, we find that in every kingdom, into which money begins to flow in greater abundance than formerly, everything takes on a new face; labor and industry gain life; the merchant becomes more enterprising, and even the farmer follows his plough with greater alacrity and attention…
>
> To account then for this phenomenon, we must consider, that though the high price of commodities be a necessary consequence of the increase of gold and silver, yet it follows not immediately upon that increase; but some time is required before the money circulates through the whole state, and makes its effect be felt on all ranks of people. At first, no alteration is perceived; by degrees the price rises, first of one commodity, then of another; till the whole at last reaches a just proportion with the new quantity of specie which is in the kingdom. In my opinion, it is only in this interval or intermediate situation, between the acquisition of money and the rise of prices, that the increasing quantity of gold and silver is favorable to industry.
>
> (Hume, *Of Money*, 1752).

Hume's opinions on the disequilibrium path of adjustment serve as a note of caution against total reliance on the comparative static results of the neoclassical model and against the belief that the pre-1936 classical economists *assumed* that the economy always functions

in full employment. Hume's analysis of disequilibrium shows that the economy can be in disequilibrium for some time and that money will not be neutral during this period.

Almost two centuries later, Pigou, a twentieth-century economist in the classical tradition, expressed similar ideas in the following excerpts from his book, *Money, A Veil* (1941).

> Money – the institution of money – is an extremely valuable social instrument, making a large contribution to economic welfare. … if there were no generally accepted money, many of these transactions would not be worth undertaking, and as a direct consequence the division of labor would be hampered and less services and goods would be produced. Thus, not only would real income be allocated less satisfactorily, from the standpoint of economic welfare, among different sorts of goods, but it would also contain smaller amounts of many, if not of all sorts. … Obviously then money is not merely a veil or a garment or a wrapper. Like the laws of property and contract, it constitutes at the least a very useful lubricant, enabling the economic machine to function continuously and smoothly. …
>
> So far everyone would be agreed. But now an important distinction must be drawn. The institution of money is, as we have seen, a powerful instrument promoting wealth and welfare. But the number of units of money embodied in that instrument is, in general, of no significance. It is all one whether the garment, or the veil, is thick or thin. I do not mean, of course, that it is immaterial whether the number of units of money is held constant, or is variable in one manner, or is variable in another manner in relation to other economic happenings. I mean that if, other things being equal, over a series of months or years the stock of money contains successively $m \times 1$, $m \times 2$, $m \times 3$. … units, it makes no difference what the value of m is. A doubled value of m throughout means simply doubled prices throughout of every type of goods – subject, of course, to the rate of interest not being reckoned for this purpose as a price – and all real happenings are exactly what they would have been with a value of m half as large. The reason for this is that, money being only useful because it exchanges for other things, a larger quantity does not, as with other things, carry more satisfaction on its back than a smaller quantity, but the same satisfaction.
>
> (Pigou, 1941, Ch. 4).

In this quote, Pigou is clearly expressing the long-run equilibrium results following an increase in the money supply. Note that what is missing in this story are Hume's conclusions on the impact of changes in the money supply during the ensuing adjustment period. However, as his other writings show, Pigou was quite aware of such an adjustment period and of the impact of money supply variations in causing fluctuations in employment and output.

14.8 The classical and neoclassical tradition, economic liberalism and *laissez faire*

The long-run equilibrium analysis of the neoclassical model in this chapter implies that the economy functions at full employment, with full-employment output, so that there is no scope for monetary and fiscal demand management policies for such an economy. Such a viewpoint is part of the classical philosophy of economic liberalism, which can be broadly formulated as stating that the economy performs at its best by itself and that the state cannot improve on its performance. This is usually also supplemented by the proposition that any intervention by the state, even with the intention of improving upon the performance of the economy,

worsens its performance. These propositions imply that the goods and input markets should be free and that free enterprise should be the desired standard. However, market imperfections such as imperfect competition, oligopoly, monopoly or monopsony, externalities, etc., could and often do exist in the actual economy. Advocates of the *strong form of economic liberalism* argue that, even in such cases, the economy should be left as it is and the state should not attempt to eliminate such imperfections; the imperfections are minor and, even when they are not of minor significance, there is no guarantee that state intervention will achieve a net improvement since its intervention might eliminate some imperfections while introducing others. A *weaker version of economic liberalism* allows the state to intervene to eliminate market imperfections through selective policies, though without assigning a role to general monetary and fiscal policies.

To be credible, the general liberalism philosophy has its underpinnings in the nation's political, economic and social ideology, and in the public's perception and goals for the actual economic and social performance of the nation. In its general approach, the underlying philosophical basis of liberalism was provided by the utilitarianism approach of Jeremy Bentham and his followers in the first half of the nineteenth century. The main tenet of this approach was that economic agents (households and firms), working in their own best interests (utility and profit maximization), would ensure that social welfare was maximized.[13] Consequently, the economy should be left alone by the government and regulatory agencies. This policy approach was summarized in the term *laissez faire*. In its economic aspects, the liberalism philosophy needed a theoretical economic model that could justify its economic policy recommendations. This model was provided at the macroeconomic level by the traditional classical approach in the pre-Keynesian period and is currently the neoclassical one – with the modern and new classical models among its versions.

The economic and social problems of nineteenth-century Britain, with rapid industrialization and urbanization, were sufficiently acute and transparent to lead to a gradual evolution of political and economic thought away from liberalism and *laissez faire* and towards some form of socialism, with support for some degree of state intervention in the economy. This evolution of ideas was widespread during the latter half of the nineteenth century and early twentieth century. The Great Depression of the 1930s destroyed the public's and economists' faith in *laissez faire*, so that Keynes's publication of *The General Theory* in 1936, with its encouragement to the state to use monetary and fiscal policies to improve on a poorly performing economy, proved to be timely and readily won acceptance from most economists and the public. Economic liberalism was eclipsed by Keynesianism for the decades from the 1930s to the 1970s. The Keynesian approach is the subject of the next chapter.

In economics, the traditional classical ideas were reformulated and rebottled in the form of the neoclassical theory during the decades of the 1940s to the 1970s. Since the 1970s, these ideas, in the form of the modern classical model and with the agenda of making microeconomics the foundation of macroeconomics, have again become the dominant approach in macroeconomics. Their return to this dominance detoured briefly through Monetarism during the 1970s. They are currently supported by the new classical approach developed during the 1970s and the 1980s.

13 The high rates of poverty, even among the "working poor," and of unemployment in Britain during the heyday of *laissez faire* economics led the nineteenth-century social reformer, Thomas Carlyle, to label economics as the "dismal science." This name for economics still remains as one of its nicknames.

14.8.1 Some major misconceptions about traditional classical and neoclassical approaches

A common misconception nowadays is that the traditional classical and neoclassical economists believed that the economy functioned well enough to maintain full employment most of the time or that it had a fast tendency to return to full employment following a disturbance and a decline in employment. In fact, many believed that "the economic system is essentially unstable" (Patinkin, 1969, p. 50).[14] Another misconception nowadays is that the classical and neoclassical economists believed that money was neutral in practice and in theory. In fact, as the business cycle literature of the nineteenth and early twentieth centuries amply shows, it was a common and strongly held belief that fluctuations in the money supply were a major cause of recessions, with declines in employment and output, and of booms in real economic activity.

During the period of traditional classical dominance (mid-eighteenth century to 1936), booms and recessions were common and sometimes quite severe. It was a common observation among economists that the velocity of circulation of money did change, and did so over booms and recessions. In fact, many economists believed that there could occur – and did occur – "extreme alternations of hoarding and dishoarding" (p. 50) because of changes in expectations, and that the changes in the money supply and in its velocity were major sources of business fluctuations, as attested by Don Patinkin (1972). Further, many economists believed that these fluctuations were:

> exacerbated by the "perverse" behavior of the banking system, which expands credit in booms and contracts it in depressions.
>
> (Patinkin, 1969, p. 51).

Among the reasons for the real effects of money supply and velocity changes was that:

> Costs have a tendency to move more slowly than do the more flexible selling prices [i.e. firms' costs were sticky relative to their final prices]
>
> (Patinkin, 1969, p. 57).

> Sticky prices are peculiarly resistant to downward pressure. … [To sum up, it was generally recognized that] cycles and depressions [are] an inherent feature of "capitalism." Such a system must use money, and the circulation of money is not a phenomenon which naturally tends to establish and maintain an equilibrium level. Its equilibrium is vague and highly unstable.
>
> (Patinkin, 1969, pp. 63–64).

In view of such strong effects of money supply variations on employment and output, what the traditional classical economists believed was that money was neutral in the long run, but not in the short run or over the business cycle. Its neutrality in the long run was often less a

14 The various quotes are taken from Patinkin (1969), even though many of them are from passages quoted by him from other writers such as Henry Simons, Frank Knight and other economists at the University of Chicago in the second quarter of the twentieth century and presumably part of the Chicago tradition of classical economics. Patinkin (1972) provides another source of the points made in these paragraphs. Parkin (1986) provides an evaluation of Patinkin's views.

matter of analysis than of belief, which was sometimes reflected in the proposition known as Say's law (see Chapter 18).

On policy issues, in view of the reasons given above, the traditional classical school believed that:

> The government has an obligation to undertake a contracyclical policy. The guiding principle of this policy is to change M so as to offset changes in V, and thus generate the full employment level of aggregate demand MV.
>
> Once a deflation has gotten under way, in large modern economy, there is no significant limit which the decline of prices cannot exceed, if the central government fails to use its fiscal powers generously and deliberately.
>
> (Patinkin, 1969, pp. 51, 63).

As the preceding quotes from Patinkin, a foremost neoclassical economist, convincingly show, monetary policy was often envisaged and recommended as a stabilization tool. Fiscal policy was sometimes, but not commonly, considered a possibility since pre-1936 economic analysis did not have its analytical basis nor did it have a theory of the aggregate demand for commodities. The analytical basis for fiscal policy and its recommendation as a major tool for stabilization of aggregate demand were due to Keynes and the Keynesians. As a counter-reformation, Barro's Ricardian equivalence theorem (Barro, 1974; see also Chapter 13, Section 13.7) sought to again remove fiscal policy from the set of potential stabilization tools.

14.9 Uncertainty and expectations in the classical paradigm

The analysis of the neoclassical model so far shows that many of its implications are contradicted by the stylized facts. In particular, monetary and fiscal policies do change output and unemployment, as against the model's implication that they do not. Therefore, the model has to be modified. Classical economists do so by introducing uncertainty into the model and relying on errors in prices expectations. The two models in this stream are the Friedman model, which relies upon wage contracts and errors in price expectations in labor markets, and the Lucas model, which relies on expectational errors by firms in commodity markets. The following analyses deal with the nature of uncertainty and such errors.

The nature of risk, uncertainty and expectations in economics

Events whose outcomes are not known at the moment of decision making used to be classified in economics in the first half of the twentieth century into risky ones or uncertain ones. The difference between these terms was that an event involves risk if the objective probabilities of its outcomes exist and are known, whereas an event involves uncertainty if the objective probabilities of its outcomes do not exist or are not known. Because of the nature of economic events and/or because of the pervasive imperfection of knowledge involving future outcomes, few economic decisions involve known objective probabilities, so that the standard case in economics is one of uncertainty. However, probability theory finds it unmanageable to include many of the elements of uncertainty such as the vagueness, inadequacy and imperfection of information that distinguish uncertainty from risk. As a consequence, neoclassical economics often abandons the above distinction between risk and uncertainty, and treats the latter as if it were really a case of risk, while Keynesian, especially post-Keynesian, economics often

makes a strong distinction between them. For consistency with the literature relevant to this chapter, we shall use the word *uncertainty* as if it were synonymous with *risk*.

Note that for uncertain events, it can be validly postulated that the individual forms subjective probabilities of the anticipated outcomes, with such probabilities being based on whatever knowledge the individual possesses or considers profitable to acquire. Such knowledge can be highly inadequate and imperfect and even the range of anticipated outcomes can differ from the possible ones, so that the subjective probabilities held can be highly erroneous or volatile[15] and would also differ among individuals. In general, classical macroeconomics ignores these problems with subjective probabilities.

Expectations hypotheses in macroeconomics

The two major hypotheses in economics for constructing the expected value of a variable are the adaptive expectations hypothesis and the rational expectations hypothesis (REH). The former is a statistical procedure, while the latter represents an economic theory of expectations. These approaches were used in Chapter 8 to estimate expected and permanent income for the demand for money function. The hypothesis used in this chapter is that of rational expectations, estimating the expected rate of inflation, the expected money supply or the expected level of aggregate demand. The material in Chapter 8 on the rational expectations hypothesis should be reviewed at this stage.

Since the classical macroeconomic models assume that the economy will either stay in equilibrium or soon revert to it, the rationally expected levels of output, unemployment and prices, as of all other endogenous variables, are their long-run equilibrium (full-employment) levels. Their values can therefore be derived from the long-run solution of the model. While this might be acceptable for analytical models that have their focus on the equilibrium values of the variables, the practice of monetary policy requires forecasts of the actual values of these variables at the time the policy will impact on them. Given the long lags in monetary policy, the relevant expectations even by the policy makers are often erroneous. From the perspective of nominal wage contracts, the economic agents must be able to forecast the price level during the duration of the contract. Such forecasts often have errors, so that the real wage received usually differs from the real wage expected by firms and workers to accrue from the contract. We next investigate the impact of expectations on wages, employment and output.

14.10 Expectations and the labor market: the expectations-augmented Phillips curve

14.10.1 Output and employment in the context of nominal wage contracts

In industrialized economies, the nominal wage between a firm and its workers is established – whether explicitly negotiated or arrived at by implicit arrangement – ahead of the production and employment decisions by the firm and before the actual price level is known. In arriving

15 Keynes (1936) emphasized the volatility of expectations in financial markets and, therefore, of the volatility of the speculative demand for money which is based on these expectations. This volatility, in turn, affected the impact of monetary policy on the economy. The treatment of subjective probabilities as if they were objective ones tends to ignore the vagueness and volatility of the former and misses their distinctive aspects and their impact on the economy. Keynes and post-Keynesians believe this impact to be important; neoclassical and modern classical economists ignore it.

at the contracted nominal wage, firms and workers must base their agreement on the *expected* real wage – that is, the nominal wage divided by the expected price level – rather than the a priori unknown *actual* real wage which will apply at the time of employment and production. However, firms can continue to adjust their employment as the price level changes, so that their decisions on employment, as determined by their demand function for labor, will depend on the actual real wage – equal to the established nominal wage divided by the actual price level. This section modifies the preceding neoclassical labor market analyses to incorporate these ideas.

To start, consider the household utility maximization, as in Chapter 3, but now with the addition of uncertainty about the future price level and with labor forming expectations on it. Let the household's expected price level for the period ahead be P^{eh}. Utility maximization would then imply that the supply function of labor is:

$$\begin{aligned} n^{\text{s}} &= n^{\text{s}}(w^{\text{eh}}) \\ &= n^{\text{s}}(W/P^{\text{eh}}) \quad n^{\text{s}\prime} > 0 \end{aligned} \tag{29}$$

where:

n^{s} = labor supply function
w^{eh} = expected real wage, as expected by labor
W = actual nominal wage
P^{eh} = price level expected by households or workers for the duration of the labor contract

Designate the inverse of $n^{\text{s}}(.)$ as $h(n)$, so that inverting (29) and rearranging yields:

$$W^{\text{d}} = P^{\text{eh}} \cdot h(n^{\text{s}}) \quad h' > 0 \tag{30}$$

where W^{d} is the wage demanded by workers in negotiations. Designate the representative firm producing the ith product – and being in the ith market – as the ith firm. Prior to the production period, the profit-maximizing ith firm in perfect competition would equate its marginal product of labor to the expected real wage measured in terms of the firm's expected product price, so that:

$$\begin{aligned} n^{\text{d}}_i &= n^{\text{d}}_i(w^{\text{ef}}_i) \\ &= n^{\text{d}}_i(W/p^{\text{ef}}_i) \quad n^{\text{d}}_{i'} < 0 \end{aligned} \tag{31}$$

where:

w^{ef}_i = expected real wage, based on the ith firm's expectations of its product price
p^{ef}_i = expected product price, as expected by the ith firm.

Aggregating over all firms, let P^{ef} be the average expected price for all firms and n^{d} the aggregate demand for labor. The aggregate demand for labor is given by:

$$n^{\text{d}} = n^{\text{d}}(W/P^{\text{ef}}) \quad n^{\text{d}\prime} < 0 \tag{32}$$

Designating the inverse of $n^{\text{d}}(.)$ as $f(n^{\text{d}})$, the nominal wage W^{o} offered in the wage negotiations by the firms is:

$$W^{\text{o}} = P^{\text{ef}} \cdot f(n^{\text{d}}) \quad f' < 0 \tag{33}$$

Assuming that the market-clearing (that is, with $n^{d} = n^{s} = n$) nominal wage is negotiated, the wage process based on (30) and (33) would yield the equilibrium nominal wage W^{c} as:

$$W^{c} = P^{ef} \cdot h(n) = P^{ef} \cdot f(n) \tag{33'}$$

where the superscript c indicates the contractual nature of the wage.[16] W^{c} can be obtained by directly solving (29) and (32) and has the general functional form:

$$W^{c} = g(P^{ef}, P^{eh}) \quad \partial g/\partial P^{ef}, \partial g/\partial P^{eh} > 0 \tag{34}$$

The explanation for the signs of the derivatives is as follows. An increase in the firm's expected price level increases its willingness to agree to higher nominal wages, and an increase in the household's expected price level makes workers demand higher nominal wages, so that a higher nominal wage will be set in the wage contracts.[17] It is further assumed that this W^{c} is set for the duration of the labor contract and that workers will supply any amount of labor demanded by firms at W^{c}. That is, for the duration of the wage contract, the *ex ante* labor supply curve is to be temporarily ignored in the analysis and the ostensible labor supply is horizontal, in the neighborhood of the equilibrium, at W^{c} in the (W, n) space.

While the firms negotiate the nominal wage on the basis of their expected price level, profit maximization by the ith firm implies, as shown by (29), that its employment and production decision depends only on its own expected price p^{ef}_{i} rather than on the price level P or the expected price level, so that its employment depends upon W^{c} and p_{i}. During the production process, the ith firm would know the actual price of its own product as a joint element of its production and pricing decision, so that actual employment will be based on the actual prices of the firms' products, rather than on the prices that had previously been expected by the firms. The average of the former is the actual price level, so that the actual aggregate employment n by firms is based on the actual real wage w. In the wage contracts context, this actual real wage is given by the contracted nominal wage W^{c} divided by the actual price level P.[18] That is, the short-run equilibrium level of employments n^{*} is given by:

$$n^{*} = n^{d} = n^{d}(W^{c}/P) \quad n^{d\prime} < 0 \tag{35}$$

Since W^{c} depends upon P^{ef} and P^{eh}, we have:

$$n^{*} = \Theta(W^{c}(P^{ef}, P^{eh})/P) \tag{36}$$

where $\partial n^{*}/\partial P > 0, \partial n^{*}/\partial P^{ef} < 0$ and $\partial n^{*}/\partial P^{eh} < 0$. The explanation for these signs is as follows. As discussed earlier, increases in the firm's expected price level or/and the household's expected price level establish a higher contractual nominal wage during wage negotiations. This, *ceteris paribus*, – i.e. without an accompanying change in the price level – increases the real wage, which reduces employment. But, for the given

16 Note that while the wage contract fixes the negotiated nominal wage, it does not fix the amount of employment, which the firm remains free to choose.

17 W^{c} will be homogeneous of degree one in P^{ef} and P^{ef}, but not in only one of them.

18 Hence, there are two stages to the process being considered. In the first stage, the nominal wage is established; in the second stage, occurring at a somewhat later date, the firm chooses the employment level on the basis of the established nominal wage and the actual selling price of its product.

contractual nominal wage, an increase in the actual price level lowers the real wage and raises employment. However, Θ is homogeneous of degree zero in P^{ef}, P^{ef} and P, so that a proportionate increase in the expected and actual price levels will not change employment, even though the nominal wage will rise proportionately.

Employment will thus depend upon the duration of the wage contract, upon the expected price levels by firms and households during wage negotiations, and upon the actual price level when employment occurs.

From (36) and the production function $y = y(n)$, with $y_n > 0$, we have the short-run equilibrium level of employment y^* as:

$$y^* = \phi(P^{ef}, P^{eh}, P) \tag{37}$$

where $\partial y^*/\partial P > 0, \partial y^*/\partial P^{ef} < 0$ and $\partial y^*/\partial P^{ef} < 0$. Therefore, for a given contractual nominal wage conditional on expectations,

$$\partial n^*/\partial P > 0,\ \partial y^*/\partial P > 0$$

For most of the commonly used forms of the production and labor supply functions, both n^* and y are homogeneous of degree zero in P^{ef}, P^{eh} and P.

Diagrammatic analysis

Figure 14.7a presents the labor demand $n^d(W/P^{ef})$ and the labor supply $n^s(W/P^{eh})$ curves. Note that the vertical axis in this figure is the nominal wage rate W. The negotiated nominal wage will be set at the equilibrium level $W^c{}_0$, and has the expected employment level of n_0^{*e}. An increase in P^{ef} will shift the labor demand curve to the right and a rise in P^{eh} will shift the labor supply curve to the left, so that each will raise the nominal wage. However, the former will increase the expected employment level and the latter will decrease it. If both P^{ef} and P^{eh} increase proportionately, the two curves will shift proportionately and the nominal wage will increase in the same proportion, without a change in the expected employment level.

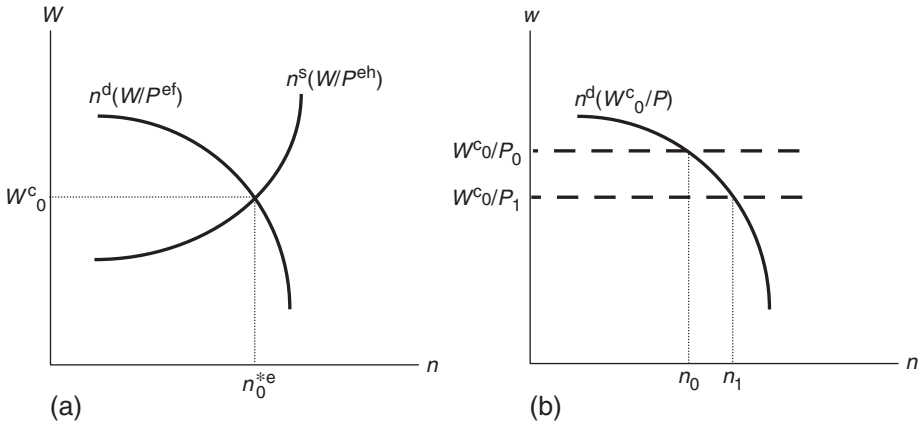


Figure 14.7

Actual employment is determined not in Figure 14.7a but in Figure 14.7b, now with the actual real wage w, equal to W/P, – in the vertical axis. For the contracted nominal wage ${W^c}_0$ from Figure 14.7a, and a given price level P_0, employment is n_0. With the contracted nominal wage still at ${W^c}_0$, a higher price P_1, $P_1 > P_0$, will lower the actual real value of the contracted nominal wage to ${W^c}_0/P_1$ and increase employment to n_1 with $n_1 > n_0$. The implicit labor supply curve is horizontal at ${W^c}_0/P$ in this figure.

If there are no errors in expectations – that is, if $P^{ef} = P^{eh} = P$, actual employment n^* will equal ${n^{*e}}_0$, as determined in Figure 14.7a, so that we can take this to be the full-employment level n^f or the "expectational (long-run) equilibrium' level n^{LR}. If P is higher than both P^{ef} and P^{eh}, $n^* > {n^{*e}}_0$, and vice versa. Therefore, the deviation in employment from its expected level n^{*e} is positively related to the errors $(P - P^e)$ in expectations.

Errors in price expectations, the duration of the wage contract and cost-of-living clauses

This deviation in employment n^* from its expected level n^{*e} will occur only during the duration of the wage contract, since the past errors in expectations will be eliminated when the wage contract is renegotiated. This is usually done through the "catch-up" cost of living clauses in labor contracts. Therefore, continuously new errors will be needed to maintain employment above n^{*e}. While this can occur for some time – the "people can be fooled some of the time" syndrome – it cannot continue indefinitely – "they cannot be fooled all the time." The former usually has to take the form of accelerating inflation rates. The latter usually occurs in two ways. One is for the future expectations of inflation to "jump" beyond the past – experienced – inflation rates in an attempt to capture the potential future acceleration in inflation. The other is to reduce or eliminate the loss in purchasing power through inflation by reducing the duration of wage contracts or by building cost-of-living clauses in them. Therefore, while the errors in expectations can induce increases in employment – and do so during accelerating inflation – such increases can only be short term and not a long-term phenomenon in practice. In particular, these increases cannot be relied upon to occur over lengthy periods or persistently high inflation rates.

A simple illustration with linear functions

Assume that the labor demand and supply functions at the time of wage negotiations are:

$$n^s = b_1 W/P^{eh} \quad b_1 > 0 \tag{38}$$

$$n^d = a_0 - a_1 W/P^{ef} \quad a_0, a_1 > 0 \tag{38'}$$

In equilibrium,

$$a_0 - a_1 W/P^{ef} = b_1 W/P^{eh} \tag{39}$$

so that the contractual nominal wage will be:

$$W^c = a_0 \left[\frac{P^{ef} P^{eh}}{a_1 P^{eh} + b_1 P^{ef}} \right] \tag{40}$$

Hence, $\partial W^{c}/\partial P^{eh}$, $\partial W^{c}/\partial P^{ef} > 0$ and a proportionate increase in both expectations increases the nominal wage rate in the same proportion. The expected equilibrium level of employment obtained by substituting this equation in the labor supply function is given by:

$$n^{*e} = a_0 b_1 \left[\frac{P^{ef}}{a_1 P^{eh} + b_1 P^{ef}} \right] \tag{41}$$

so that a proportionate increase in both expectations does not change n^{*e}. Note that the employment level is not set in the wage contract and can deviate from n^{*e}.

At the time of production, with the nominal wage set by the wage contract at W^{c}, the actual real wage and employment will be:

$$w^* = a_0 \left[\frac{P^{ef} P^{eh}}{a_1 P^{eh} + b_1 P^{ef}} \right] \left(\frac{1}{P} \right) \tag{42}$$

$$n^* = n^{d} = a_0 - a_1 a_0 \left[\frac{P^{eh} P^{ef}}{a_1 P^{eh} + b_1 P^{ef}} \right] \left(\frac{1}{P} \right) \tag{43}$$

Both w^* and n^* are homogeneous of degree zero in P, P^{ef} and P^{eh}. If P exceeds both of its expectations, the real wage will turn out to be less than its expectation in the wage contract and n will be greater than n^{e} – and vice versa. If there are no errors in expectations, $P = P^{ef} = P^{eh}$, so that $n = n^{f} = a_0 - a_0 a_1/(a_1 + b_1)$. This expectational equilibrium level of employment is independent of the price level and is therefore the classical full-employment level. Note that positive (negative) expectational errors can induce actual employment to be less (greater) than this full-employment level.

14.10.2 The Friedman supply rule

The preceding types of analyses often assume that:

$$P^{ef} = P \tag{44}$$

This assumption is usually justified by the argument that, for profit maximization, *each* firm only needs to know the price of its own product – in which it possesses a great deal of information – and its own factor costs, represented by the contractual nominal wage, but does not need to know the prices of all the commodities in the economy. Since the average price level P in (42) is only a proxy for the average of the individual commodity prices, each of which is set by the firm supplying the commodity, the firms on average can be expected to predict P fairly accurately.

By comparison, utility optimization by a household requires knowledge of the general price level in order to calculate the purchasing power of the nominal wage. To know the price level requires knowledge of all the commodity prices, which is a degree of knowledge that each household rarely, if ever, possesses. Hence, it is assumed that households individually and on average in the aggregate cannot predict the price level with sufficient accuracy, so that P^{eh} can differ from P.

With these assumptions, the employment and output supply functions can be restated in the more specific form:

$$n^* = \theta(P/P^{eh}) \quad n^{*\prime} > 0, \partial n^*/\partial P > 0, \partial n^*/\partial P^{eh} < 0 \tag{45}$$

$$y^* = \phi(P/P^{eh}) \quad y^{*\prime} > 0, \partial y^*/\partial P > 0, \partial y^*/\partial P^{eh} < 0 \tag{45'}$$

If $P > P^{eh}$, $w < w^e$, so that labor will prove to be unexpectedly cheaper and firms will employ more than they had expected to employ. Hence, $y^{*\prime} > 0$, $\partial y^*/\partial P > 0$ and $\partial y^*/\partial P^{eh} < 0$. (45) and (45′) are homogeneous of degree zero in P and P^{eh}.

14.10.3 Expectations-augmented employment and output functions

Since the expectations of both firms and households are negatively related to output y, we can simplify the notation by replacing them by the single variable P^e. Therefore, the short-run equilibrium output function becomes:

$$y^* = y(P/P^e) \quad \partial y/\partial P > 0, \partial y/\partial P^e < 0 \tag{46}$$

where P^e is now the expected price level for both firms and households.

There will not be any errors in expectations when $P^e = P$. Designate the respective levels of employment and output when there are no errors in households' price expectations as n^{LR} (or n^f) and y^{LR} (or y^f), consistent with the earlier long-run analysis of this chapter in which there was perfect foresight so that there were no errors in price expectations. These employment and output values are therefore the long-run equilibrium levels of employment and output.

The *log-linear* form of (46) is:

$$\ln y^* = \ln y^{LR} + \beta(\ln P - \ln P^e) \quad \beta > 0 \tag{47}$$

(47) is the *expectations-augmented output function* or *Friedman's supply function*. Note, again, the difference between the superscript LR, which designates the full-employment level (without errors in expectations), and *, which designates the short-run equilibrium level in the presence of errors in expectations. Correspondingly, we have for the level of employment:

$$\ln n^* = \ln n^{LR} + \alpha(\ln P - \ln P^e) \quad \alpha > 0 \tag{48}$$

(48) is the *expectations-augmented employment function*.

Note that the price expectations in these equations refer to those incorporated in wage contracts. The economy deviates from its full-employment level due to errors in these expectations; compared with the full-employment level, output is greater if $P > P^e$ and lower if $P < P^e$. In the former case, real wages are lower than the full-employment real wage, making it attractive to hire more labor than in the error-free equilibrium; while the opposite holds in the latter case.

Define *expectational equilibrium* as the state where there are no errors in expectations. That is, it requires that:

$$P^e = P \tag{49}$$

From (47) and (48), in the long-run expectational equilibrium:

$$n = n^{\mathrm{LR}} = n^{\mathrm{f}} \tag{50}$$

$$y = y^{\mathrm{LR}} = y^{\mathrm{f}} \tag{51}$$

which assert that in expectational equilibrium the levels of employment and output are the full-employment levels and are independent of the price level. (47) and (48) assert that deviations from these levels occur because of expectational errors, with positive errors ($P > P^{\mathrm{e}}$) causing an increase in output and employment, and negative errors ($P < P^{\mathrm{e}}$) causing a decrease.

14.10.4 The short-run equilibrium unemployment rate and Friedman's expectations-augmented Phillips curve

Unemployment U equals $(L - n)$ and the unemployment rate u equals U/L $(= 1 - n/L)$. Therefore, the approximation for the relationship between the log values of u^*, L and n^* is:

$$\ln u^* = \ln L - \ln n^* \tag{52}$$

From (48) and (52),

$$\ln u^* = \ln u^{\mathrm{n}} - \alpha(\ln P - \ln P^{\mathrm{e}}) \quad \alpha > 0 \tag{53}$$

where $\ln u^{\mathrm{n}}$ $(=\ln u^{\mathrm{LR}} = \ln L - \ln n^{\mathrm{LR}})$ is the natural rate of unemployment. u^* is the short-run equilibrium unemployment rate in the presence of errors in expectations. (53) is the *expectations-augmented Phillips curve* (EAPC), proposed independently by Friedman (1968) and Phelps (1968) as a correction of the Phillips curve. In a dynamic context, P is replaced by π (the inflation rate) in the Phillips curve equation, so that the usual form of the EAPC is stated as:

$$\ln u^* = \ln u^{\mathrm{n}} - \alpha(\ln \pi - \ln \pi^{\mathrm{e}}) \quad \alpha > 0$$

Note that several assumptions were needed for deriving this equation. Among these was the assumption of market clearance in the labor and commodity markets, an assumption that many Keynesians would not accept. Also note that (53) places the burden of all possible deviations from the natural rate of unemployment on errors in price-level expectations over the duration of the nominal wage contract.

Implications of the expectations-augmented Phillips curve for fluctuations in unemployment

The preceding arguments have the following three implications:

(a) Sources of deviations of the unemployment rate from the natural rate, other than those due to errors in expectations, are ruled out by the EAPC. In particular, the numerous sources of deviations considered by the Keynesians and new Keynesians are not captured.[19]

19 Among these sources of deviations are: lack of labor market clearance, coordination failures among goods markets, market inefficiencies such as adjustment lags, market imperfections such as firm and labor immobility, etc.

Among these is the failure of the labor market to clear on a continuous basis, such as in recessions, as well as the possibility of persistent underemployment equilibria in certain rare circumstances such as in the Great Depression of the 1930s. For the new Keynesians, it is the failure to include market imperfections, such as monopolistically competitive firms, that is important (see Chapter 15).

(b) *If* the expectational errors are insignificant in magnitude or are not relevant because of a very short duration of labor contracts,[20] any fluctuations in unemployment over time would have to be explained by a theory of fluctuations in the natural rate of unemployment.

(c) In the EAPC, $u^* < u^n$ requires $w^* < w^{LR}$ and $P^* > P^{LR}$, and vice versa, so that fluctuations in employment and output occur because of changes in price level (inflation) but not because of those changes in real aggregate demand which are not reflected in prices (inflation).

Given the experience of considerable cross section and business cycle variations in the unemployment rates in real-world economies, classical theory supplements its theory of expectational errors with a theory of fluctuations in the natural rate of unemployment over time and across countries. In particular, if this theory is to explain the experienced rates of unemployment, it cannot assume that the natural rate is a constant, so that an explanation of the long-run variations in the unemployment rate will require a long-run theory of changes in the natural rate, while an explanation of the cyclical variations in unemployment will require a theory of the cyclical variations in the natural rate.

These arguments lead to two competing sets of theories of unemployment. These are:

1 The classical theories, with fluctuations in the natural rate itself both over the business cycle and the long run, accompanied by continuous labor market clearance but with expectational errors to explain deviations of the actual rate of unemployment from the natural rate. Changes in aggregate demand do not cause fluctuations in unemployment unless they cause prior unanticipated changes in the price level.

2 The Keynesian theories of unemployment, which also allow – but do not require – changes in the natural rate, but emphasize deviations of the actual rate of unemployment from the natural rate due to fluctuations in the demand for commodities and labor, especially in the presence of market imperfections such as contractual rigidities, sticky prices, efficiency wages, etc. Chapter 15 expands on these factors.

Empirical validity of the Friedman supply hypothesis based on errors in price expectations in labor markets

The Friedman output model implies that anticipated monetary policy would not change output and unemployment, which is contradicted by the stylized facts given at the beginning

20 In periods of high and volatile inflation rates, labor tries to protect itself against positive errors in expectations by shortening the duration of the wage contract. Thus, during the 1950s with relative price stability, labor contracts of three years' duration were not uncommon. In the inflationary 1980s, three- and two-year contracts became rarer and one-year contracts or contracts with cost of living adjustments in the nominal wage became much more common. The result of such shortening of the contract duration was to reduce the impact of expectational errors and make the expectations-augmented Phillips curve shift to a more vertical position.

of this chapter. Further, the Friedman model asserts that the effects of monetary policy changes, both anticipated and unanticipated, must go through errors in the price expectations embedded in wage contracts and a subsequent decrease in real wages. This too is contradicted by the stylized facts: an expansionary monetary policy increases output and reduces unemployment without necessarily producing a prior change in the price level or a decrease in real wages. Therefore, Friedman's short-run model of commodity supply does not provide a satisfactory explanation of the short-run impact of monetary policy on output and unemployment.

14.11 Price expectations and commodity markets: the Lucas supply function

The EAPC is based on the possibility that unanticipated price changes generate expectational errors in real wages in the labor market. This emphasis on the labor market is in some ways more consistent with the Keynesian aggregative analysis, though the expectations-augmented Phillips curve is associated with Milton Friedman. Neoclassical analysis focuses more on the markets for commodities and its microeconomic Walrasian basis, as in Chapter 3, which implies that the output of each commodity depends upon its relative price. Lucas (1972, 1973) and Sargent and Wallace (1975) modified the certainty version of the microeconomic model by introducing into it uncertainty and firms' expectations on relative product prices. This section presents one version of the Lucas analysis.[21]

Lucas supply function or rule

Assume as in the preceding section that firm i produces good X_i, with the quantity produced as x_i and sold at the price p_i. The firm buys inputs, of which labor is taken to be the only variable input. Given the nominal wage W, profit maximization by the firm in perfect competition implies that its supply function is given by:

$$x_i = x_i(p_i/W) \quad x_i' > 0 \tag{54}$$

with p_i determined in the perfectly competitive market i. While the firm is not directly concerned with the price level P, Lucas's analysis assumes that the nominal wages change proportionately with the price level, so that P can be used as an index of labor cost.[22] Therefore, replacing W by P,

$$x_i = x_i(p_i/P) \quad x_i' > 0 \tag{55}$$

21 See Walsh (2003, Ch. 5) for another version of the Lucas model.

22 For the price level to serve as an index of labor costs, the nominal wage rate must instantly adjust proportionately to the actual price level. Hence, there must be nominal wage flexibility and instant adjustment in nominal wages, without any rigidities introduced by labor contracts or money illusion in the labor supply function. Compare this with the expectations-augmented Phillips' curve, also in the neoclassical tradition, which assumes that nominal wages do not rise proportionately with the actual price level but do so only with respect to the expected price level, so that real wages do change because of the unanticipated part of inflation.

For simplification, Lucas assumes that the price p_i in market i deviates in percentage terms from P by an amount z_i which is normally distributed, independent of P and has a zero expected value and variance η^2. Therefore,

$$p_i = P + z_i \tag{56}$$

Hence, z_i $(= p_i - P)$ defines the deviation of the ith product's price from the price level. The product price p_i is called the "local price" while z_i $(= p_i/P)$ is referred to as the "relative price,' so that the change in the local price of the ith product incorporates changes in both the price level and its relative price. Both increases in the general price level and in the firms' relative product price can occur at any time. The ith firm is assumed to know the price of its own product but not to know the price level, which it estimates conditional on the information available to it. Given such uncertainty, re-specify (55) as:

$$x_{it} = x_i(p_{it}/E(P^e{}_t|I_t(i)) \quad x_i{}' > 0 \tag{57}$$

where:

x_{it} = output in market i in period t
$E(P^e{}_t|I_t(i)$ = mean (mathematical expectation) of the price level expected for period t by firms in market i, conditional on $I_t(i)$
$I_t(i)$ = information available in market i in period t.

Specify the log-linear form of (57) as:

$$x_{it} = x^*{}_{it} + \gamma[p_{it} - E(P^e{}_t|I_t(i))] \quad \gamma > 0 \tag{58}$$

where *all the variables are now in logs* and $x^*{}_{it}$ is the ith firm's output under perfect certainty or if there are no expectational errors. γ is the firm's response to an increase in its relative price.

Lucas (1972, 1973) provides a specific procedure for determining the expected relative prices. Firms use the available information on aggregate demand and supply movements, and on local and general prices, to form their expectations on the distribution of local and general prices in the present period.[23] This provides them, at the beginning of the current period, with a prior distribution of the expected price level P^e, with mean $\underline{P}$ and constant variance σ^2, with this distribution formed prior to the observation of the current local prices.[24] During period t, the firm knows $\underline{P}$, σ^2, η^2 and observes p_i. The (ith) firm uses this knowledge of its local price to calculate $E(P^e{}_t|I_t(i))$ as:

$$E(P^e{}_t|I_t(i)) = \underline{P}_t + [\sigma^2/(\eta^2 + \sigma^2)][p_{it} - \underline{P}_t] \tag{59}$$

where:

$\underline{P}_t$ = mean of the prior distribution of expected prices
σ^2 = expected variance of P ("price level variability")
η^2 = expected variance of z_i ("relative price variability")
$\sigma^2 + \eta^2$ = expected variance of p_i ("local price variability").

23 This would include knowledge of past shifts in demand and supply and past variations in local and general prices, but can also include any available information about the future.

24 These assumptions make the deviation of the relative price a random one.

The second term on the right of (59) is the correction made to the prior expectation $\underline{P}_t$ of the price level on the basis of the observed local price p_{it}.

Equation (59) was justified on the basis of information available in the ith market when the price level is not directly observed. This view of the nature of information was couched by Lucas (1972) in what has come to be known as the *island parable*. This parable envisions the workers and firms as distributed spatially over islands (or isolated points). The firms do not know about activity (prices and output) on other islands but must forecast the average price level (over all the islands) in order to formulate their labor demand and output supply decisions. To forecast the price level, they use the historic variability of their island price – represented by $(\eta^2 + \sigma^2)$ – relative to overall variability to forecast the shift of the price level from a prior expected level, and do so as specified in (59).

Rewrite (59) as:

$$E(P^{e}{}_t | I_t(i)) = \alpha \underline{P}_t + (1 - \alpha) p_{it} \tag{60}$$

where $\alpha = \eta^2/(\eta^2 + \sigma^2) \geq 0$, so that α is the expected ratio of the relative price variance to the total local price variance.

Substituting (60) in (58), we have:

$$x_{it} = x^{*}{}_{it} + \alpha\gamma[p_{it} - \underline{P}_t] \tag{61}$$

Integrating (61) over all markets i, with total output supply designated as y^{s}, and replacing the known local price p_{it} by the *actual* aggregate price level P_t,

$$y^{s}{}_t = y^{*}{}_t + \alpha\gamma[P_t - \underline{P}_t] \tag{62}$$

which is the *aggregate supply function* based on firms' expectations about variations in local and general prices. The two components of a firm's response to an expectation error are: (a) γ, which is the change in output in response to the expectation error $(P - P^{e})$; and (b) α, which is the revision of P^{e} from its prior value $\underline{P}$. (62) is known as the *Lucas supply function or rule*.

If $\eta = 0$ and $\alpha = 0$, so that relative prices are expected to be stable, (62) becomes:

$$y^{s}{}_t = y^{f}{}_t \tag{63}$$

In this case with $\eta = 0$, aggregate supply will not respond to absolute price changes and hence to changes in aggregate demand, with the result that the aggregate supply function will be vertical in the (y, P) space. But if $\sigma \to 0$ – that is, the price level is expected to be stable, so that the change in local price is taken to be wholly a relative price change – the aggregate supply function will become:

$$y^{s}{}_t = y^{*}{}_t + \gamma[P_t - \underline{P}_t]$$

where γ is likely to be positive since it reflects firms' responses to an increase in their relative prices.

Another explanation of the Lucas supply function

Equation (62) was based on price misperceptions. It is to be distinguished from a somewhat similar equation, also attributed to Lucas, which can be derived from the

intertemporal substitution of work in household decisions. In this model of intertemporal utility maximization, for given nominal wages in each period and a given current price level P_t, if the expected price level (P^e_{t+1}) rises, it will decrease the expected future real wage (w^e_{t+1}) for the given nominal wage,[25] while the current real wage (w_t) is unaffected. Hence, from utility optimization, workers will substitute work in the current period (i.e. increase n^s_t by decreasing their leisure time in t) for work in future periods (i.e. decrease n^s_{t+1} by increasing leisure in $t+1$). Conversely, they would work – and produce – less in the present period if, at the given current price level, the expected future price is lower and the expected real wage higher. Such behavior would produce recessions in output in the latter case and booms in the former, thus causing real business cycles. However, the empirical significance of such a model is limited since the observed intertemporal substitution of work in labor supply decisions on the basis of price movements is too low to imply the larger observed variations in output over the business cycle, so that we shall ignore it further in this chapter.

Comparing the Friedman and the Lucas supply functions

We are, therefore, left with two types of expectations-based Phillips curve relationships. One of these is Friedman's expectations-augmented Phillips curve, which is based on errors in expectations in labor market and contractual rigidities; and the other is Lucas's supply function, based on errors in the expectations of relative prices in commodity markets. The former is sometimes claimed to be an example of the latter under the argument that nominal wages are the "local" price of workers as suppliers of labor and the observed increases in these nominal wages are used by the workers in forming their expectations on the price level and the real wage (the "relative price" of labor). However, the theoretical and empirical bases of the two are quite different,[26] so that it is preferable to keep them separate for analytical reasons. But they are similar in spirit in that both maintain full employment unless there are errors in expectations. Note that the errors in price expectations will be corrected as time passes since this will provide information on the actual prices. In modern economies, the lag in information on actual prices and inflation rates is often a month or a few months, so that any expectational errors would be corrected fairly soon. Therefore, both the errors in expectations and the deviation of short-run equilibrium output from the full-employment output will be *self-correcting* and *transient*.

14.12 The Lucas model with supply and demand functions

The Lucas supply function derived above is:

$$y^s_t = y^*_t + \alpha\gamma[P_t - \underline{P}_t] \tag{62}$$

On the *aggregate demand function*, Lucas (1972, 1973) assumed that:

$$Y^d_t = Y^d_{t-1} + \delta + \mu_t \tag{64}$$

25 Note that this assumption of the rigidity of the future nominal wage in the face of the expected rise in the price level seems unrealistic.

26 In particular, the expectations-augmented Phillips curve requires contractual rigidities in nominal wages while the Lucas supply rule does not have such contractual rigidities. In the EAPC, if there is an unanticipated rise in prices, workers' real wage falls, so that it is the workers who lose. In the Lucas model, the unanticipated rise in prices fools firms and their profits fall.

where:

Y^{d} = nominal aggregate expenditures/demand
δ = systematic (known) increase in demand
μ = random shift in demand, with $E\mu = 0$

Further, from the definition of nominal expenditures and noting that all variables are in logs:

$$Y_t = y_t + P_t \tag{65}$$

Assuming equilibrium in the commodity market with $y^{\mathrm{d}} = y^{\mathrm{s}} = y$ and $Y^{\mathrm{d}} = Y^{\mathrm{s}} = Y$, eliminate y_t from (62) to (65). Then, assuming rational expectations as applied by the classical paradigm (which is that the expected level of a variable is its long-run equilibrium level), $Ey_t = y^{\mathrm{LR}}{}_t$ and $\underline{P} = EP_t$, we get:

$$P_t = \frac{\alpha\gamma\delta}{1+\alpha\gamma} + \frac{1}{1+\alpha\gamma}\, Y_t + \frac{\alpha\gamma}{1+\alpha\gamma}\, Y_{t-1} - y^{\mathrm{LR}}{}_t \quad \alpha \geq 0,\ \gamma > 0 \tag{66}$$

Starting from (65), substitute (64) for Y_t, and (66) for P_t.. This yields:

$$y^*{}_t = y^{\mathrm{LR}}{}_t - \frac{\alpha\gamma\delta}{1+\alpha\gamma} + \frac{\alpha\gamma}{1+\alpha\gamma}(Y_t - Y_{t-1}) \quad \alpha \geq 0,\ \gamma > 0 \tag{67}$$

Given (64), (67) in equilibrium (with $y^{\mathrm{d}} = y^{\mathrm{s}} = y$) simplifies to:

$$y^*{}_t = y^{\mathrm{LR}}{}_t + \frac{\alpha\gamma}{1+\alpha\gamma}\,\mu_t \quad \alpha \geq 0,\ \gamma > 0 \tag{68}$$

This equation forms the theory of output of the modern classical school. It shows that the modern classical school does not assume full employment or maintain its continuous existence. It does assert that deviations in output from the full-employment level can be caused by errors in price expectations; however, any such deviations will be self-correcting and transient under rational expectations. The causes of the deviations of actual output from short-run equilibrium are ruled out by the way the analysis is formulated. In this theory, the only fluctuations in output that are not transitory and self-correcting have to come from fluctuations in full-employment output due to shifts in technology and physical capital, labor supply and human capital, and availability of resources. Real business cycle theory rests on the preceding Lucas model and expands on this theme.

If firms believe that all the variation in prices is in the general price level and none in the relative price level, $\eta = 0$, so that $\alpha = 0$ and (68) yields the error-free output as $y^{\mathrm{LR}}{}_t$. That is:

$$y^*{}_t = y^{\mathrm{LR}}{}_t \text{ for } \eta = 0$$

so that $\partial y_t/\partial\delta = \partial y_t/\partial\mu = 0$. Therefore, (68) implies that, if firms do not perceive any changes in relative prices, both systematic and random shifts in aggregate demand will not cause deviations in output from its full-employment level under certainty. But if relative prices are expected to change, $\eta > 0$ and >0. In this case, random – but not systematic – shifts in aggregate demand can have real effects, since from (68),

$$\partial y^*{}_t/\partial\delta = 0 \quad \text{for} \quad \eta > 0 \tag{69}$$

while:

$$\frac{\partial y^*_t}{\partial \mu_t} = \frac{\alpha\gamma}{1+\alpha\gamma} = \frac{\eta^2\gamma}{\sigma^2+\eta^2(1+\gamma)} > 0 \text{ for } \eta > 0 \tag{70}$$

Note from (70) that even the *random* shifts in aggregate demand cause changes in output *only* if firms misinterpret its impact on the price level and believe that some part of the resulting increase in prices is a relative price increase. In conditions of hyperinflation where the likelihood and magnitude of general price increases dominate over relative price increases, the public's expectation is likely to be $\eta = 0$, so that even random changes in aggregate demand will not have any effects on output. In this context, neither systematic nor random demand increases will change output. Therefore, for hyperinflations, money is likely to be neutral for both systematic and random increases in the money supply. Hence, even random increases in the money supply will not always induce increases in output and employment.

Asymmetric information and the impact of systematic demand increases on output

Equation (69) states that systematic demand increases by the policy makers will not change real output. But, suppose that the policy makers systematically (for example, through the use of a rule such as the Taylor rule) increase demand but in such a way (as by a one-time shift in the coefficients of the rule followed) that the firms interpret it to be a random increase. This is a case of asymmetric information between the policy maker and the public. This "disguised or misinterpreted" nature of systematic demand increase will allow the policy maker to increase real output through the "random multiplier" $\partial y/\partial \mu$, whose value is specified by (70). However, the systematic nature of the demand increase will, sooner or later, be observed by firms, leading to two types of change. One, firms will find it optimal to acquire better information so that their likelihood of correctly perceiving a systematic demand increase as being systematic – rather than erroneously as random – will increase. For this correctly perceived systematic demand increase, there will be no change in output. Second, firms faced with repeated increases in the general price level will modify their expectations on prices by increasing the value of σ^2 relative to η^2, so that $\partial y/\partial \mu$ will decrease. In the limiting case, if there were only systematically induced increases in the price level, firms would adjust their expectations such that $\eta^2/(\sigma^2+\eta^2) \to 0$, so that $\partial y/\partial \mu$ will go to zero. That is, systematic demand increases, disguised or misinterpreted as being random ones, will eventually lose their efficacy.

The implications of the Lucas model for unemployment

To examine the response of unemployment to changes in aggregate demand, start with the definition of unemployment as being $(L - n)$ and use the production function $y = y(n)$, $y'(n) > 0$, to go from y to u. Then, (68) implies that employment n is a function of the random demand term μ_t but not of the systematic demand term δ. With all variables being in logs, and assuming a log-linear relationship, (68) implies that:

$$u^*_t = u^{\text{n}} - \beta\mu_t \quad \beta \geq 0 \tag{71}$$

where u^* is the short-run equilibrium unemployment rate, β is a positive function of $[\alpha\gamma/(1+\alpha\gamma)]$ with $\beta = 0$ if $\alpha = 0$. First note that δ is not in (71), just as it was not in

(68), so that $\partial u_t^* / \partial \delta = 0$. Second, note that α is in (71) through β, just as it was in (68), and that $\beta > 0$ for $\alpha > 0$ and $\gamma > 0$. For $\beta > 0$, unemployment decreases if there is a random increase in aggregate demand, since the latter would persuade individual firms that the relative demand and the relative price of their product have increased. That is, for $\beta > 0$, (71) implies a negatively sloped curve in the (y, u) space, which looks like a Phillips curve – but is not the Phillips curve[27] – and can be called the "Lucas–Phillips curve." Hence, there is a seeming tradeoff – for *given* expectations on general versus relative price increases – between price level increases and output increases in the short run.

However, there is a vital difference between (71) and the standard Phillips curve (for which, see Chapter 15). While the latter was envisaged by the Keynesians as a durable tradeoff between price increases (both anticipated and unanticipated) and unemployment, (71) cannot be used as a durable tradeoff for policy purposes. Meaningful policy increases in demand cannot be random but must be systematic, such as by a constant δ in (64) above or according to some established rule, and, under the Lucas analysis, any systematic demand changes cannot change real output. Further, as argued above, any systematic demand or price increases, disguised or misinterpreted as being random ones, will sooner or later lose their efficacy as their systematic nature becomes understood.

Hence, according to (71), correctly anticipated inflation – such as would be the case if the inflation rate were constant or steadily increasing – has a vertical Lucas–Phillips curve, thereby not providing a durable impact of money, prices and inflation on output. This is now generally accepted. However, it is now also accepted that in the short run, monetary policy, whether operating through money supply or interest rates, does affect output and does so earlier than prices and inflation, as indicated in the stylized facts at the beginning of this chapter. It is now generally accepted that these facts cannot be explained in a satisfactory manner by imperfect information, resulting in temporary price or inflation misperceptions, but must rest on theoretical foundations and empirical factors not encompassed in Lucas's analysis.

Empirical validity of the Lucas supply model based on price misperceptions in the commodity market

The Lucas output model (Lucas, 1972, 1973; Sargent and Wallace, 1976; see also Chapter 17) implies that anticipated monetary policy would not change output and unemployment. This clearly contradicts the stylized facts given at the beginning of this chapter. Further, the Lucas model asserts that the effects of monetary policy changes, both anticipated and unanticipated, must go through price level changes and errors in price expectations. This too is contradicted by the stylized facts: an expansionary monetary policy increases output and reduces unemployment without necessarily producing a prior change in the price level, as Lucas (1996) concluded later from his assessment of the empirical evidence (see Section 14.16). Therefore, for the short run, the Lucas model does

27 It is not the Phillips curve or the EAPC because these are based on real wages actually falling as inflation occurs, while the Lucas version does not allow this fall. His curve emanates from the firm's belief that relative prices are increasing when they are not, so that the perceived inflation rate is below the actual one. Since, under our assumptions, nominal wages increase proportionately with the actual rate of inflation, the firm's relative price misperceptions imply that the firm's perception is that the real wage is falling, even though it does not do so. This has the consequence that the firm's real profits fall even as it supplies more output – so that its output increase is likely to be short lived.

not provide a satisfactory explanation of the impact of monetary policy on output and unemployment.

14.13 Defining and demarcating the models of the classical paradigm

Evolution of the classical paradigm

The principles of the classical paradigm evolved out of several rather disparate elements. The dominant one was the economic and political philosophy of liberalism in the first half of the nineteenth century. Its economic analysis was that of individual markets for commodities, factors of production, money and bonds, with competition as the invisible hand guiding each of them to equilibrium through the adjustment of the relevant price. In this analysis, all prices were flexible and adjusted to bring each market into equilibrium, so that each market cleared at the trading price. Much of this analysis of individual markets was set out in the following rather separate categories: individual commodity and labor markets to determine relative prices, wages, output and employment; quantity theory for determining the price level and the loanable funds theory for determining the interest rate; and business cycle theories. These distinctive theories nowhere appeared in an integrated format so that one cannot point to the work of any economist prior to Keynes's book *The General Theory* (1936) for a statement of an integrated macroeconomic theory or model. In particular, there was no integrated model that included the consumption and saving functions and the investment multiplier, and therefore, no aggregative theory of the commodity market. Further, while there was a great deal of discussion about the nature of risk and uncertainty, the above components of the traditional classical approach did not incorporate it in a meaningful way. These are the elements of what Keynes labeled in Chapter 1 of *The General Theory* as the classical model, though there did not exist in the literature at that time such an established and complete macroeconomic model. We have labeled this pre-Keynesian model the *traditional classical model*, in comparison with the subsequent neoclassical and modern classical models.

Keynes's *The General Theory* for the first time provided an integrated macro model of the economy and also fundamentally altered the way the profession models short-run aggregative economics. He chose to give the foremost place in his model to the commodity market, with an analysis incorporating the multiplier and also money demand analysis. John Hicks (1937) organized Keynes's ideas on the commodity and monetary sectors into what he labeled the IS–LM framework, and cast the traditional classical model in the same format to facilitate comparison. The traditional classical model thus recast in the mould of the IS–LM framework, which had been proposed to illuminate Keynes's ideas, came to be known as the neoclassical model. The IS–LM model of aggregate demand, combined with the AD–AS model for the determination of output and the price level, provided the first formally integrated macroeconomic model of the classical paradigm. It was elaborated and refined in the decades up to 1970.

Neoclassical economics with the addition of rational expectations, if there is uncertainty, and an insistence on continuous labor market clearance has been labeled in this book the *modern classical model*.[28] The combination of the modern classical model with Ricardian

28 Note there are many significant differences between the somewhat diffuse ideas of the nineteenth century and the modern classical model. Among these differences is the explicit commodity market analysis incorporating

equivalence has been labeled the *new classical model*. In general, these models of the classical paradigm do not incorporate market imperfections, that is, deviations from the perfect markets assumption, which are emphasized in Keynesian economics (see next chapter).

There can be considerable disputes about the proper delineation of the classical schools or models. We introduced the following taxonomy in Chapter 1. We elaborate on it here, though at the risk of some repetition. No claim is being made to this taxonomy being a universal – or perhaps even a majority – one. We have chosen it below for reasons of clarity in separating each model from the others, rather than leaving their differences ambiguous, while maintaining consistency with the writings and folklore in the history of economics thought.

All the schools of the classical paradigm share the common belief that the real-world economy under consideration – and not just the models – functions at full employment in the long run and that one of the characteristics of long-run equilibrium is the independence of the real variables from the financial ones, so that money is neutral in the such equilibrium. Further, all schools share the belief that deviations from the long-run equilibrium can occur in the short run but such deviations are self-correcting and transient. States with less than full employment are, therefore, states of disequilibrium during which the economy continues to adjust towards its full-employment equilibrium – and not away from it, A major difference among these schools is whether the real-world economy adjusts so fast as to have continuous equilibrium, so that it will not show any evidence of disequilibrium, even though disequilibrium remains a hypothetical state within the model.

The traditional (pre-Keynesian) classical ideas

This section lays out our interpretation and distillation of the writings of the pre-Keynesian economists. Their ideas were not expressed or formulated in terms of a compact model, and the analysis of the expenditure sector (the IS curve) and the multiplier was not available to them. Their common belief was that output and employment in the long run equilibrium depended upon the real sector's relationships only and were independent of the monetary sector. The economy's interest rate was determined by the theory of loanable funds, which in modern terminology corresponds to the market for bonds (see Chapter 19). Further, the quantity theory for the determination of the price level applied in equilibrium. But outside this equilibrium, changes in the money supply could change output and employment. Not only could the money supply affect the real sector in this way, the economy was considered to be very prone to fluctuations in output and employment. Many of these fluctuations were attributed to money supply shocks or the response of the money supply to real shocks. In particular, most of the classical economists did not believe that the economy functioned so well that it always maintained full employment or that it did so most of the time. In fact, recessions and crises – many of them originating in the banking sector or financial speculation or occurring due to the response pattern of the financial sector to real shocks – were common, and widely recognized as such, during the nineteenth century. Hence, the traditional classical school did not assume continuous full employment or that it existed most of the time.

the multiplier concept, the interest sensitivity of money demand, the elimination of the direct transmission mechanism of money supply effects on nominal income, explicit analysis of government deficits and their implied future taxes, rational expectations, etc.

The neoclassical model

This model was an attempt to bottle the main ideas of the pre-Keynesian classical economists into a compact modern macroeconomic model. This process was initiated by Hicks in 1937, who borrowed the IS–LM analysis from Keynes's work and used it as a technique for interpreting the traditional classical ideas, thereby re-incarnating those ideas into the neoclassical model. The neoclassical model continued to have both equilibrium and disequilibrium aspects and did not assume instantaneous market clearance. In this, it represents the ideas of the pre-Keynesian classical economists more faithfully than does the modern classical model.[29]

The modern classical and new classical models

The certainty version of the modern classical model modifies the neoclassical model by adding the assumption of continuous market clearance, especially of the labor market at the full-employment level. By doing so, it ignores the disequilibrium properties and multipliers of the neoclassical model as being irrelevant in practice. The uncertainty version of this model adds in the rational expectations hypothesis. Some economists would also add to this mix the assumption of Ricardian equivalence. However, our definition of the modern classical model excludes this assumption, only making it part of the new classical model. This differentiation means that, under our definitions, fiscal policy would change aggregate demand in the modern classical model but not in the new classical model.

Hence, under our designations, the constituents of the *modern classical model* are:

1 the neoclassical model, modified by the additions of:
2 uncertainty, with deviations of output and employment from their long-run values due to errors in price expectations;
3 the rational expectations hypothesis, which implies that the errors in expectations will be random;
4 continuous market clearance (especially of the commodity and labor markets).

The constituents of the *new classical model* are:

1 the modern *classical* model, modified by the addition of:
2 Ricardian equivalence.

Note that both the modern classical and the new classical models do possess money supply changes as a policy tool for changing aggregate demand in the economy, but, of the two, only the modern classical model allows fiscal policy to change aggregate demand. For the long run, both these models imply the neutrality of money in the full-employment state, so that the impact of the money supply and velocity changes can only be on the price level and not on real output and employment. Further, for the short run, unanticipated changes in money supply and velocity can cause deviations of short-run output and employment

29 This modern classical model is also different from the traditional classical model in other ways. One of these relates to the existence of speculative balances in the modern classical model, whereas the quantity theory component of the traditional classical model did not have such balances.

from their long-run values, but, given rational expectations, these deviations will be transient and self-correcting as new information on prices becomes available. Therefore, both the modern classical and the new classical models imply that there is neither a need nor scope for systematic monetary policy for changing the levels of output and employment in the economy, so that such policies should not be pursued. These ideas are further explored in Chapters 15 and 17.

14.14 Real business cycle theory and monetary policy

Business cycles are cyclical fluctuations in the economy's output and employment in real, not analytical, time. Their explanation relates to the short term, which is a chronological concept of time, rather than the analytical short run or long run.

Real business cycle theory is an offshoot of the modern classical model and asserts that business fluctuations occur *only* in response to shocks to the fundamental determinants of long-run output and employment (e.g. see Prescott, 1986; Christiano and Eichenbaum, 1992; Romer, 1996, Ch. 4). These determinants are technology, which determines the production function and the demand for inputs, and the supply of factor inputs. Among the determinants of the latter are preferences, including those on labor supply, which depends on the labor–leisure choice and the stock of resources. Shifts in the production function or input supplies alter long-run equilibrium output, as well as being a source of cyclical fluctuations in output. The real business cycle theory derives the fundamental determinants of business cycles from the general macroeconomic models of the classical paradigm.

Explicitly, or by omission, real business cycle theory also holds that shifts in aggregate demand, no matter what their source, do not cause changes in output and employment and therefore do not cause business cycle fluctuations. Therefore, changes in consumption, investment, exports, money supply and demand (or the central bank's interest rate policy) or fiscal deficits cannot change output and employment. This exclusionary proposition is derived from the properties of the long-run equilibrium of the modern classical model. To be valid, it requires perfectly competitive markets and also that long-run equilibrium is *continuously* maintained in the economy.

The policy implication of real business cycle theory, as of the modern classical model of which it is an elaboration, is that systematic monetary (and fiscal) policies cannot affect output and employment, so that they cannot be used to moderate the business cycle. The critical elements for this implication are the Friedman–Lucas supply equation and rational expectations, according to which anticipated changes in prices, inflation and monetary policy cannot affect output. Therefore, the Taylor rule, under which systematic monetary policy manipulates aggregate demand by changing the interest rate in response to the output gap and the deviation of inflation from its target rate, can only be useful in controlling inflation but not in moderating the output gap. According to the modern classical school, while random monetary policy can change aggregate demand, the central bank cannot predict and therefore cannot offset the random fluctuations in the private components of aggregate demand. In short, in the new classical model, monetary policy and the Taylor rule have no legitimate role in moderating or reducing the duration of business cycles.

Intuitively, the problem with the real business cycle theory is most evident in its explanation of recessions. It attributes recessions to a fall in labor productivity and/or an increase in the preference for leisure. The objections to these explanations are succinctly stated by the quip: recessions occur because "workers forget how to do things" ("lose some of their knowledge") and/or because they decide to become lazier for some time, thereby causing the recessionary

fall in output! Neither of these explanations is plausible, so the validity of the real business cycle theory is highly doubtful. Looking at upturns in business cycles, the real business cycle theory attributes upturns to increases in productivity and/or increases in the preference for work over leisure. The latter is hardly plausible over the length of upturns in the economy, while the former is highly plausible. Here, however, it is the plausibility of the assertion of real business cycle theory that aggregate demand increases cannot also be a source of upturns that is highly doubtful.

The real business cycle propositions rest on the assumption that all markets can be taken to be competitive and efficient (i.e. continuous equilibrium) in the economy. This assumption is not consistent with models of the Keynesian paradigm, since they incorporate market imperfections and/or failure of the economy to achieve long-run equilibrium instantly after a demand shock. In these models, shifts in aggregate demand, whether through shifts in investment and other private sector variables or in monetary and fiscal policy, can produce changes in output and be a source of, or contribute to the continuation of, business cycles. More specifically on monetary policy, market imperfections can create non-neutrality of money, so that fluctuations in the money supply can add to output fluctuations. Conversely, the appropriate monetary policy can reduce the severity of cyclical fluctuations due to aggregate demand shocks coming from the private sector. Further, Keynesians do not deny that shifts in the fundamental determinants of output, mentioned above, can also cause output fluctuations.

Therefore, the core of the debate about the validity of real business cycle theory is not about whether shocks to technology and factor inputs can cause cyclical fluctuations, for that is not in dispute. It is rather about whether shocks to aggregate demand can cause such fluctuations and whether monetary policy can moderate them. Real business cycles and the modern classical school deny that they can, or do so in a significant manner, while Keynesians assert that they can do so. This issue is easily testable by the appropriate causality tests. The consensus on the empirical evidence seems to be that the major part (in some estimates, as large as 70 percent) of the fluctuations in output can be attributed to productivity shocks. This is a testament to the success of real business cycle theory, as compared with Keynesian ideas from the 1940s to the 1970s that had attributed most business cycle fluctuations to shifts in aggregate demand. However, the empirical evidence leaves a very significant part of the fluctuations in output that cannot be explained by shifts in technology and preferences. Overall, the empirical evidence, as well as intuition, seems to indicate that fluctuations in aggregate demand, in addition to changes in technology and preferences, do cause fluctuations in output and employment and that money supply growth is positively related to output growth. Therefore, real business cycle theory is not strictly valid, and monetary policy can be pursued in appropriate cases to reduce output fluctuations.

The exponents of the real business cycle theory also prefer to test this theory by the calibration and simulation of models rather than by the econometric testing of their hypotheses. The former procedure requires a priori specification of the likely values of the parameters, on which there can be considerable doubts. Further, the findings may not be robust to small changes in these assumed values, or consistency with the empirical observations may require implausible values. Consequently, this testing procedure and its reported findings have not won general acceptance.

There seem to be at least two major contributions of the real business cycle theory. One, it has firmly established that changes in technology and preferences do cause cyclical fluctuations in output and may do so significantly more than fluctuations in aggregate demand. Two, the approach initiated by the real business cycle agenda to macroeconomic modeling is now firmly established. This approach requires that macroeconomics be based on

optimization over time by individual economic agents in a dynamic context. This stochastic dynamic intertemporal approach to macroeconomics permeates current macroeconomic models, including the new Keynesian model, which is presented in the next chapter. The major deficiency and unrealistic assertion of the real business cycle theory is that it denies demand shifts any role in output fluctuations.

The empirical evidence on the impact of changes in aggregate demand on output is often on the impact of money supply changes, which change aggregate demand, on output. The influential study by Friedman and Schwartz (1963a,b) used evidence from over 100 years of US data to show clear evidence that money supply changes lead, and therefore Granger-cause, changes in real economic activity. However, inside money (i.e. deposits in banks) is the largest component of money. Subsequent contributions by other authors showed that deposits respond to macroeconomic disturbances, so that money is more highly correlated with lagged output than with future output; i.e. deposits lag rather than lead output. However, monetary aggregates such as M2 still lead output. Further, if the central bank uses the interest rate as its operating monetary policy target, and money supply responds endogenously to it, the evidence seems to show that changes in interest rates lead output.

To conclude, empirical evidence shows that while shocks to real factors such as technology and preferences do cause fluctuations in output, shocks to monetary policy variables of money and interest rates also do so. Models of the modern classical school and real business cycle theory do not provide a satisfactory explanation for the latter finding. In recent years, sticky price and inflation models of the new Keynesian school have been proposed to explain economic fluctuations. An example of these studies is provided by Ireland (2001b).

14.15 Milton Friedman and monetarism

Milton Friedman occupies a special place in the counter-reformation from Keynesian economics to the neoclassical and eventually to the modern classical theories, though his ideas are, in many ways, closer to the neoclassical economics of the 1960s and 1970s than to the modern thinking. In the 1950s, Friedman argued and showed through his theoretical and empirical contributions that "money matters" – that is, changes in the money supply change both nominal and real output – is against the then general view of the Keynesians that changes in the money supply brought about through monetary policy did not significantly affect the economy, or did so unpredictably.[30] He argued and tried to establish through empirical studies that, as far as nominal national income was concerned, the money-income multiplier was more stable than the investment-income multiplier, so that monetary policy was more predictable than fiscal policy in its impact on nominal national income. However, Friedman held that major instability in the US economy had been produced or, at the very least, greatly intensified by monetary instability and that major depressions were associated with monetary contractions, prior to and after the establishment of the Federal Reserve System in 1913. Therefore:

> The first and most important lesson that history teaches about what monetary policy can do … is that monetary policy can prevent itself from being a major source of economic disturbance. [However, while] monetary policy can contribute to offsetting major disturbances in the economic system arising from other sources … we simply

30 See Chapter 2 for the elucidation of his 1956 article.

> do not know enough to be able to recognize minor disturbances when they occur or to be able to predict what their effects will be with any precision … [so that monetary policy should only offset major] disturbances when they offer "a clear and present danger.'
>
> (Friedman, 1968, pp. 12–14).

Friedman is famous for his assertion that inflation is always and everywhere a monetary phenomenon and that increases in the money supply will produce inflation, not increases in real output, in the long run. Another aspect of Friedman's agenda to re-establish the doctrine that money "matters" for short-run fluctuations in output and employment was to set out in the 1950s and 1960s the theory – and to establish empirically – that money demand was a function of a few variables and that the money demand function was stable, with the result that the velocity of money also had a stable function. We have already discussed some of these contributions in Chapter 2 in the context of Friedman's restatement of the quantity theory of money. These arguments had been accepted by the profession by the early 1960s, and contributed to the conversion of Keynesian macroeconomics to a Keynesian–neoclassical synthesis expressed by the IS–LM model for the determination of aggregate demand.

On the relationship between the nominal variables and the real side of the economy, Keynesians in the late 1950s and 1960s had relied on the Phillips curve, which showed a negative tradeoff between the rate of inflation and the rate of unemployment. Friedman argued that the natural rate of unemployment – and, therefore, full-employment output – was independent of the anticipated rate of inflation, so that the fluctuations in output and the rate of unemployment were related to deviations in the inflation rate from its anticipated level. This relationship came to be known as Friedman's expectations-augmented Phillips curve and incorporated his contributions on the natural rate of unemployment.

While Friedman brought the role of anticipations on the rate of inflation into discussions on the role and effectiveness of monetary policy in the economy, he did not use the theory of rational expectations; the rational expectations hypothesis had not yet entered the literature and Friedman relied on adaptive expectations in his empirical studies.

Hence, Friedman was a precursor of the modern classical school but not fully a member of it. Nor does this school follow all of his ideas. He was closer to the Keynesians in one important respect than to the later modern classical school. He believed, as did the Keynesians, that the economy does not always maintain full employment and full-employment output – and does not always function at the natural rate of unemployment, even though this concept was central to his analysis. Hence, policy-induced changes in aggregate demand could induce short-term changes in output and employment. Therefore, money mattered even to the extent that changes in it could induce changes in employment and output, depending upon the particular stage of the business cycle. While this view was shared with the Keynesians, Friedman tilted against the Keynesians on the pursuit of discretionary monetary policy as a stabilization tool – especially for "fine tuning" the economy – because of his belief that the impact of money supply changes on nominal income had a *long and variable* lag. He reported on the lag in the impact of monetary policy that:

> The rate of change of the money supply shows well-marked cycles that match closely those in economic activity in general and precede the latter by a long interval. On the average, the rate of change of the money supply has reached its peak nearly 16 months before the peak in general business and has reached its trough over … 12 months before the trough in general business. … Moreover, the timing varies considerably from cycle to cycle – since 1907 the shortest time span by which the money peak preceded the

> business peak was 13 months, the longest 24 months; the corresponding range at troughs is 5 months to 21 months.
>
> (Friedman, 1958; see Friedman, 1969, p. 180).

With such a long and variable lag between changes in the money supply and nominal income, the monetary authorities cannot be sure when a policy-induced increase in the money supply would have its impact on the economy. Such an increase in a recession may not, in fact, increase aggregate demand until the following boom, thereby only increasing the rate of inflation at that time. Consequently, Friedman argued that discretionary monetary policy, intended to stabilize the economy, could turn out to be destabilizing. Friedman's recommendation on monetary policy was, therefore, that it should maintain a low constant rate of growth, as stated by him in:

> There is little to be said in theory for the rule that the money supply should grow at a constant rate. The case for it is entirely that it would work in practice. There are persuasive theoretical grounds for desiring to vary the rate of growth to offset other factors. The difficulty is that, in practice, we do not know when to do so and by how much. In practice, therefore, deviations from the simple rule have been destabilizing rather than the reverse.
>
> (Friedman, 1959).[31]

Therefore, while both Friedman and the modern classical economists are opposed to the pursuit of discretionary monetary policy, they arrive at this position for quite different reasons. For Friedman, money supply changes can change output and employment but the long and variable lags in this impact make a discretionary policy inadvisable; over time; it could make the economy perform worse rather than better. For the modern classical economists, the economy maintains full employment except for transitory and self-correcting deviations from it due to random errors in price expectations, so that systematic policy changes in the money supply cannot change output and employment, but only the price level. Further, for this school, the lags in the impact of systematic money supply changes on nominal national income are not significant.

On the transmission mechanism from money supply changes to income changes, Friedman supported Fisher's direct transmission mechanism (from money supply changes directly to expenditures) over the indirect one (from money supply to interest rates to investment in the Keynesian and IS–LM models). Neoclassical and modern classical models espouse the latter rather than the former.

Monetarists and the St Louis equation

Monetarism and monetarists have been defined in a variety of ways. In a very broad sense, monetarism is the proposition that "money matters" in the economy. In this sense, Friedman, Keynes and the Keynesians[32] were all monetarists, while the modern classical

31 The above prescriptions for policy were repeated in Friedman (1968).

32 This was not true of many Keynesian models popular in the 1950s and 1960s. Some of these relegated money to a minor role, since they claimed that money was only a small part of the economy's liquidity, which included trade credit, etc. Other models claimed that the price level and the inflation rate were an outcome of the struggle between

school is less monetarist since it downplays the impact of money supply changes on the real variables of the economy. In a narrow sense, monetarism as a label was associated with the St Louis school in monetary and macroeconomics. We shall define monetarism in this narrow sense. The St Louis school provided in the late 1960s and early 1970s an empirical procedure for estimating the relationship between nominal income and the money supply. This was the estimation of a reduced-form equation (Andersen and Jordan, 1968) of the form:

$$Y_t = \alpha_0 + \Sigma_i a_i M_{t-i} + \Sigma_j b_j G_{t-j} + \Sigma_s c_s Z_{t-s} + \mu_t \tag{72}$$

where:

Y = nominal national income
M = nominal value of the appropriate monetary aggregate
G = value of the appropriate fiscal variables
Z = vector of the other independent variables
μ = disturbance term.

Equation (72) is called the St Louis monetarist equation and was presented earlier in Chapter 8. While its common form used nominal income as the dependent variable, the dependent variable can be changed, depending upon the researcher's interest, to real output, the unemployment rate, the rate of inflation or some other endogenous variable. In general, the St Louis equation is a reduced-form estimation equation of the short-run macro models, with the monetary aggregates and the fiscal variables being taken as exogenous.

The St Louis equation has become a popular method for determining the impact of monetary and fiscal policies on nominal national income and other variables. Its initial estimation by researchers at the Federal Reserve Bank of St Louis (Andersen and Jordan, 1968) showed that the money aggregates had a strong, positive and rapid impact on nominal income, this impact being more significant than that of fiscal policy. The marginal money-income multiplier was about 5 over five quarters, while the marginal impact of fiscal policy was positive for the first year and then turned negative, with a multiplier of only about 0.05 over five quarters.[33] These findings were consistent with Friedman's stance, except that the estimations of the St Louis equation indicated a much shorter and more reliable lag than Friedman had found. Therefore, contrary to Friedman's recommendations and consistent with those of the Keynesians, monetarism was consistent with the stance that monetary policy could be useful for short-term stabilization.

The St Louis monetarism represented a transitional stage in the transition from Keynesian ascendancy in economics in the decades before 1970 to the ascendancy of the neoclassical and modern classical schools in the 1980s and 1990s. In many ways, it was an amalgam of Keynesian and Friedman's ideas in macroeconomics, and led the way to the re-emergence of the classical doctrines. While its theoretical underpinnings have not survived, its impact

firms and unions over relative income shares. Still others claimed that firms followed a full-cost pricing policy, with the money supply accommodating itself to the resulting price level because the central bank did not want the unemployment rate to rise.

33 Numerous applications of the St Louis equation showed that its empirical findings differed among countries, periods and the definitions of the policy variables. However, their basic conclusion, that money supply changes have a strong short-term impact on the economy, remained fairly robust.

on monetary policy has proved to be longer lasting. On this, its contributions were that money matters, that the control of the money supply is important for containing inflation and that the responsibility for inflation rests with the central bank.

14.16 Empirical evidence

One way of assessing the empirical validity of the implications of the classical models is by comparing their implications with the stylized facts set out early in this chapter. For the long run, the classical models imply that output and unemployment in the economy are independent of the money supply, price level and inflation, while the relationship between money supply and price level is proportionate. Both implications seem to be confirmed by data over long periods of time (Kormendi and Meguire, 1984; Geweke, 1986; McCandless and Weber, 1995; Taylor, 1996; Lucas, 1996), though some studies show a positive correlation between output and inflation in a general context of low inflation (McCandless and Weber, 1995) while others show a negative one, especially at higher inflation rates (Barro, 1996).

On the short-run or dynamic impact of an expansionary monetary policy, i.e. an expansionary money supply and/or decrease in interest rates, the modern classical model implies, through the Friedman–Lucas supply analysis, that prices and inflation will increase. If the increase is unanticipated, output will rise and unemployment will fall; if it is anticipated, there will be no change in output and unemployment. However, as the stylized facts show, the dynamic response of output and unemployment to an expansionary monetary policy is hump-shaped: output first increases (and unemployment falls) for several quarters, and then begins to decrease (Sims, 1992; Ball, 1993). This evidence contradicts the response pattern implied by the modern classical school. Further, while the impacts of anticipated and unanticipated monetary policy can sometimes be different, this difference is not always significant. In any case, an anticipated monetary policy usually does have a significant impact on output and unemployment, and unexpected price and inflation changes explain only a small fraction of the output changes that occur. Further, the real effects of monetary policy do not proceed through prior changes in prices and inflation rates, as asserted in the Friedman and Lucas supply rules (Lucas, 1996).

Therefore, the inescapable conclusion is that while the modern classical model does provide an acceptable long-run relationship between money and output/unemployment (i.e. money and monetary policy are neutral), it does not provide a satisfactory short-run theory of the impact of money and monetary policy on output and unemployment.

This chapter has presented the analyses of the classical paradigm in short-run macroeconomics. How does it perform empirically? For this, we rely on the assessment provided by Robert E. Lucas, who is associated with the modern classical school and has been a major contributor to it.

Lucas on the neutrality versus non-neutrality of money

In the "Nobel lecture" (1996), given on his receipt of the Nobel Prize in economics, Robert Lucas noted that:

> In summary, the prediction that prices respond proportionately to changes *in the long-run*, deduced by Hume in 1752 (and by many other theorists, by many different routes, since), has received ample – I would say decisive – confirmation, in data from many times and places….

> The observation that money changes induce output changes in the same direction receives confirmation in some data sets but is hard to see in others. Large-scale reductions in money growth can be associated with large-scale depressions or, if carried out in the form of a credible reform, with no depression at all.
>
> (Lucas, 1996, p. 668; italics added).

> *Sometimes*, as in the U.S. Great Depression, *reductions in money growth seem to have large effects on production and employment*. Other times, as in the ends of the post-World War I European hyperinflations, large reductions in money growth seem to have been neutral, or nearly so. Observations like these seem to imply that a theoretical framework such as the Keynes–Hicks–Modigliani IS/LM model, in which a single multiplier is applied to all money movements regardless of their source or predictability, is inadequate for practical purposes.
>
> (Lucas, 1994, p. 153; italics added).

Note that this quote states that changes in the money supply need not always be neutral and can (not must) have large real effects, lasting over a considerable period. But the modern versions of the classical paradigm do not provide a theory to explain such significant instances of the non-neutrality of money.

Lucas on the validity of the modern classical analysis for the short run

Lucas claims that "Macroeconomic models with realistic kinds of monetary non-neutralities do not yet exist." (Lucas, 1994, pp. 153–4). "… anticipated and unanticipated changes in money growth have very different effects" (1996, p. 679). However, on the models that attribute this non-neutrality to unanticipated or random changes in the price level, the evidence shows that:

> *Only small fractions of output variability can be accounted for by unexpected price movements. Though the evidence seems to show that monetary surprises have real effects, they do not seem to be transmitted through price increases, as in Lucas (1972)*.
>
> (Lucas, 1996, p. 679; italics added).

Note that this quote indicates that money supply changes do not have to first cause changes in prices before they affect real output. That is, output may change in response to monetary policy changes, even though the markets might not first or ever adjust prices. This implies that an expansionary monetary policy would produce output increases before being reflected in inflation (Mankiw, 2001). This is a powerful indictment of the long-run classical paradigm with its underlying assumption at both the microeconomic and macroeconomics levels, which is that following an increase in demand, prices are first adjusted by markets, and this is followed by adjustments in quantities demanded and supplied by economic agents.

Lucas on the state of macroeconomic theory

> Little can be said to be firmly established about the importance and nature of the real effects of monetary instability, at least for the U.S. in the postwar period. Though it is widely agreed that we need economic theories that capture the non-neutral effects of

> money in an accurate and operational way, none of the many available candidates is without serious difficulties.
>
> (Lucas, 1994, p. 153; italics added).

Hence, given the above assessment by Lucas, it is fair to conclude that the short-run modern classical models, based on his model of price misperceptions in the commodity market and/or on Friedman's errors in expectations occurring in wage contracts, fail to provide a satisfactory explanation of the empirical evidence on the impact of monetary policy on output, employment and unemployment. We must therefore continue the search for additional theories of such impact. The next chapter looks at various theories in the Keynesian paradigm, and their success in this objective.

Conclusions

The traditional classical model, dominant through most of the nineteenth and early twentieth centuries, was a somewhat disorganized set of ideas, for which we can envisage full employment as the long-run equilibrium state but with the existence of disequilibrium being a distinct possibility. Changes in the money supply affected employment and output and caused business cycle fluctuations. The main components of the traditional classical analysis were the quantity theory of prices and the theory of the rate of interest. There was no specific theory of employment and output, though there was a general belief that, after a monetary disturbance, the economy eventually returns to its full-employment level. There was also no treatment of the product market with its concept of the investment multiplier.

The neoclassical model reorganized the traditional classical analysis into a compact form, based on the IS–LM apparatus, and explicitly incorporated into it the analysis of the product market through the IS curve. It also replaced the quantity theory by a general equilibrium determination of the price level. In a departure from many other treatments of the neoclassical model, our version, presented in this chapter, emphasizes both its equilibrium – at the full-employment level – and disequilibrium analyses. In particular, it does not assume that the latter can be ignored for applications to real-world economies and in its policy recommendations.

In our definitions of the various schools that have evolved out of the neoclassical school, the addition of the assumptions of rational expectations and continuous equilibrium to the neoclassical model turns it into the modern classical model, which allows for short-run deviations from full employment due to errors in expectations. Adding further the doctrine of Ricardian equivalence to the modern classical model defines the new classical model. This latter model implies that fiscal deficits do not even affect the commodity market equilibrium or aggregate demand in the economy, so that the fiscal variables cannot be used as a policy tool. However, the money supply remains a policy tool for changing aggregate demand in the economy.

The implications drawn from the neoclassical model are usually for its full-employment equilibrium states. For such states, output is also at the full-employment level. Further, since the neoclassical economists usually assume that equilibrium is restored within a reasonable or acceptable period of time, the neoclassical prescription is that the economy should be left alone to achieve equilibrium. Monetary and fiscal policies should not be pursued. If pursued, they cannot affect the equilibrium real values of the macroeconomic variables, but would only change the nominal values. Few economists and central banks accept these

propositions for the short-run impact of monetary and fiscal policies. Most, if not all, central banks actively pursue monetary policies to modify the short-run performance of the economy.

Versions of the equilibrium part of the neoclassical model provided, in the past, the theoretical basis for the schools in political economy known as Reaganomics in the USA and Thatcherism in Britain. These philosophies of political economy advocated the cessation of any attempts by the monetary and fiscal authorities to change real output and employment in the economy, and the restriction of monetary policy solely to the goal of price stabilization. These schools are no longer popular.

The short-run version of the modern classical model, in which deviations of output from the full-employment level occur through errors in expectations due to unanticipated price inflation, is clearly not valid. This is clearly so, as a comparison of this model's implication with the stylized facts on the impact of anticipated monetary policy on output, set out at the beginning of this chapter, show. Further, according to Lucas, who is one of the major contributors to this model, "only small fractions" of the actual deviations occur due to errors in price expectations, so that the classical paradigm does not explain the major part of the actual deviations from full employment. Correspondingly, the expectations-augmented Phillips curve and the Friedman and Lucas supply rules do not explain the major part of the deviations of the rate of unemployment from the natural rate. Bluntly put, in Lucas's assessment of the empirical evidence the classical model is not a satisfactory one for explaining (i) the departures from full-employment output and from full employment of workers, (ii) the way in which monetary policy impacts on output, or (iii) the extent of that impact. These conclusions are now generally accepted in the literature, so that the Friedman–Lucas notion of price misperceptions resulting from imperfect information in a model with price and wage flexibility is no longer considered to be an appropriate explanation of the short-run non-neutrality of monetary policy. It is therefore important to provide other reasons for it. The Keynesian paradigm provides additional and distinctive explanations for this short-run non-neutrality. The next chapter sets out the Keynesian paradigm and evaluates its empirical validity.

Summary of critical conclusions

- ❖ In the long-run general equilibrium of the economy, the quantity of money is neutral and monetary policy cannot be used to increase or decrease the level of output or unemployment. It can only change the nominal value of aggregate demand, the price level and the rate of inflation.
- ❖ In the short run, the introduction of uncertainty and expectations into the neoclassical model by Friedman and Lucas produces a short-run equilibrium in which employment and output differ from the long-run levels due to errors in price expectations.
- ❖ The assessment of the empirical evidence indicates that, in the short run, output and employment do depend on monetary and fiscal policies.
- ❖ The major part of the observed impact of monetary policy does not first go through changes in prices nor occur because of errors in price expectations, as it must do in the modern classical model.
- ❖ Consequently, the models of price/inflation misperceptions are inappropriate for explaining short-run deviations in output, employment or unemployment from their long-run levels, or the short-run non-neutrality of money.

Review and discussion questions

1. "In the context of the modern classical model, the rate of money growth is the sole determinant of the trend rate of inflation, and the stance of fiscal policy is irrelevant to it." Discuss.
2. Discuss the analytical and empirical validity of the statement that the Pigou effect ensures that full-employment equilibrium exists if prices and wages are flexible.
3. Suppose that the government wants to increase its expenditure g and has the options of financing it by higher taxes, bond issues or increases in the monetary base. Further, when g is increased through bond or monetary financing, the central bank can also undertake offsetting open-market operations. What combinations, if any, of financing methods and open-market operations will allow the following goals to be met in the neoclassical model:

   (i) no change in the price level;
   (ii) no change in investment;
   (iii) no change in P and investment.

4. Present the analysis of the statement that the effects of government deficits on the rate of inflation and real output and unemployment depend on the way in which the deficit is financed.

   Also, analyze the above statement for the following cases:

   (i) The central bank is known to follow the rule of stabilizing the growth of the money supply.
   (ii) Government debt (bonds) is only sold to the central bank.
   (iii) The central bank is known to follow the rule of stabilizing the real rate of interest.

5. The monetary authority has decided to adopt one of the following money supply rules:

   (a) $M^s = kPy$
   (b) $M^s = ky$

   where $k > 0$. Show their implications for aggregate demand, prices and output in the context of (i) the neoclassical model and (ii) the neoclassical model with a zero speculative demand for money (i.e. $\partial m/\partial R = 0$). Is either one of these policies viable?
6. One of the banking innovations in the 1960s was the payment of interest on certain types of demand deposits. Assume that interest is paid on money at the nominal rate R_m, which equals $(R-x)$, where x is the nominal return on bonds, which is exogenously determined by market structures and the cost of servicing deposits.

   (i) Use Baumol's transactions demand model to derive the demand function for money.
   (ii) Generalizing the above demand function to $m^d(y, R, x)$, shows the behavior of the LM curve for shifts in x and P.
   (iii) What is the effect of an increase in x on aggregate demand, output and price level in the neoclassical model?
   (iv) Assuming that both R and x always increase by the expected rate of inflation, do (ii) and (iii) again.

7. "For the economy to have a determinate price level and money to have a positive value, it is necessary that the economy have a demand for money and a mechanism for limiting

its supply." Discuss in the context of (a) the quantity theory, (b) neoclassical economics with an exogenous money supply, and (c) privately issued monies.

8. "In a closed economy, if the money stock is held constant by the central bank, an increase in the government deficit does not have either short-run or long-run effects on aggregate demand and output." Discuss in the context of the neoclassical model and the new classical model (with rational expectations and Ricardian equivalence).
9. Discuss whether the existence of business cycles and the observed positive correlation between real and monetary variables mean that the modern classical models are neither valid nor relevant for policy purposes.
10. Specify a model that generates real business cycles only. Discuss whether this model allows for the observed cyclical correlation between money and output.
11. Present the new classical model (with rational expectations and Ricardian equivalence). Analyze the role of monetary policy in this model.

    How would this model explain the major recession in the early 1990s in most developed economies? What monetary policies – if any – are consistent with this model for moderating the effects of the recession on output and unemployment?
12. Discuss the proposition that a change in the rate of growth of the money supply will not affect output and unemployment in the short run, as well as in the long run, if wages and prices are fully flexible.
13. "The rate of money growth is the sole determinant of aggregate demand and the trend rate of inflation, and the stance of fiscal policy is irrelevant." Discuss.
14. "In response to demand shocks, short-term price adjustments by markets occur earlier than quantity adjustments at the level of both the firm and the economy." Discuss the relevant theory behind this statement. Also, discuss its empirical validity at the macroeconomic level.
15. The "crowding out" hypothesis asserts that investment is crowded out by fiscal deficits. Is it fully or partially valid in (i) the IS–LM model, (ii) the neoclassical model, or (iii) the long-run equilibrium of the modern classical approach? As part of your answer, discuss: can complete crowding out occur in the context of (a) a financially advanced economy, (b) a financially under-developed economy?
16. The 1970s monetarism and the new classical school are sometimes lumped under the same banner. How do these schools view the existence of short-run unemployment in the economy and what policies does each imply for curing it? What role do the schools assign to monetary policy in this respect?
17. What are main tenets of the modern classical school and how do they differ from those of the traditional classical school?

    Discuss critically the contributions of the modern classical school to our understanding of the role and limitations of monetary and fiscal policies for stabilization, and compare these with those of the neoclassical model.
18. Discuss the empirical validity of the implications of the modern classical model for (i) the long run, (ii) the short run.
19. What is the empirical assessment by Lucas of the importance of short-run fluctuations in unemployment and output due to errors in expectations? Discuss other possible reasons for such fluctuations.
20. How do (i) real business cycle theory, (ii) Friedman–Lucas supply analysis, explain fluctuations in unemployment over the business cycle? What is the impact on unemployment of changes in aggregate demand in these models? How would you test the validity of the implications of these theories?

21. If the central bank cannot change output and unemployment through systematic money supply changes and its induced changes in the anticipated inflation rate, should it try to do so by changes in the unanticipated inflation rate? If such unanticipated inflation requires a random change in the money supply, how can the central bank achieve this? What will be the effect of such a random change on output and unemployment, and what conclusions can be drawn on the advisability of such a monetary policy? Discuss.

References

Andersen, L.C., and Jordan, J.L. "Monetary and fiscal actions: a test of their relative importance in economic stabilization." *Federal Reserve Bank of St. Louis Review*, 50, 1968, pp. 11–24.

Ball, L. "How credible is disinflation? The historical evidence." *Federal Reserve Bank of Philadelphia Business Review*, 1993, pp. 17–28.

Barro, R.J. "Are government bonds net wealth?' *Journal of Political Economy*, 82, 1974, pp. 1095–118.

Barro, R.J. "Unanticipated money growth and unemployment in the United States." *American Economic Review*, 67, 1977, pp. 101–15.

Barro, R.J. "Inflation and growth." *Federal Reserve Bank of St. Louis Review* 78, 1996, pp. 153–69.

Brayton, F., and Tinsley, P. "A guide to FRB/US: a macroeconomic model of the United States." *Federal Reserve Board Finance and Economic Discussion Series*, 1996–42, 1996.

Bullard, J., and Keating, J.W. "The long-run relationship between inflation and output in postwar economies." *Journal of Monetary Economics*, 36, 1995, pp. 477–96.

Christiano, L.J., and Eichenbaum, M. "Current real-business-cycle theories and aggregate labour-market fluctuations." *American Economic Review*, 82, 1992, pp. 430–50.

Christiano, L.J., Eichenbaum, M., and Evans, C. "Monetary policy shocks: What have we learned and to what end." In J. Taylor and M. Woodford eds, *Handbook of Macroeconomics*, vol. 1A. Amsterdam: Elsevier North-Holland, 1999, pp. 65–148.

Cover, J.P. "Asymmetric effects of positive and negative money supply shocks." *Quarterly Journal of Economics*, 107, 1992, pp. 1261–82.

Eichenbaum, M. "Comments: interpreting the macroeconomic time series facts: the effects of monetary policy." *European Economic Review*, 36, 1992, pp. 1001–11.

Friedman, B.M. "The LM curve: a not-so-fond farewell." *NBER Working Paper* no. 10123, 2003.

Friedman, M. "The supply of money and changes in prices and output" (1958). Reprinted in M. Friedman *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine, 1969, pp. 171–87.

Friedman, M. *A Program for Monetary Stability*. New York: Fordham University Press, 1959.

Friedman, M. "The role of monetary policy." *American Economic Review*, 58, 1968, pp. 1–17. Reprinted in Milton Friedman, *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine, 1969, pp. 95–110.

Friedman, M. "A theoretical framework for monetary analysis." *Journal of Political Economy*, 78, 1970, pp. 193–238.

Friedman, M. "Nobel prize lecture: inflation and unemployment." *Journal of Political Economy*, 85, 1977, pp. 451–73.

Friedman, M., and Schwartz, A. "A monetary history of the United States, 1867–1960." Chicago: University of Chicago Press, 1963a.

Friedman, M., and Schwartz, A. "Money and business cycle." *Review of Economics and Statistics*, 45, 1963b, pp. 32–64.

Geweke, J. "The superneutrality of money in the United States: an interpretation of the evidence." *Econometrica*, 54, 1986, pp. 1–22.

Hicks, J.R. "Mr. Keynes and the classics: a suggested interpretation." *Econometrica*, 5, 1937, pp. 147–59.

Hume, D. *Of Money* (1752). Reprinted in *The Philosophical Works Of David Hume*, 4 volumes. Boston: Little, Brown, 1954.

Ireland, P.N. "Money's role in the monetary business cycle." *NBER Working Paper* no. 8115, 2001a.

Ireland, P.N. "Sticky-price models of the business cycle: specification and stability." *Journal of Monetary Economics*, 47, 2001b, pp. 3–18.

Keynes, J.M. *The General Theory of Employment, Interest and Money*. New York: Macmillan, 1936.

Kormendi, R.C., and Meguire, P.G. "Cross-regime evidence of macroeconomic rationality." *Journal of Political Economy*, 92, 1984, pp. 875–908.

Lucas, R.E., Jr. "Expectations and the neutrality of money." *Journal of Economic Theory*, 4, 1972, pp. 103–24.

Lucas, R.E., Jr. "Some international evidence on output–inflation tradeoffs." *American Economic Review*, 63, 1973, pp. 326–34.

Lucas, R.E., Jr. "Comments on Ball and Mankiw." *Carnegie-Rochester Series on Public Policy*, 41, 1994, pp. 153–5.

Lucas, R.E., Jr. "Nobel lecture: monetary neutrality." *Journal of Political Economy*, 104, 1996, pp. 661–82.

McCandless, G.T., Jr., and Weber, W.E. "Some monetary facts." *Federal Reserve Bank of Minneapolis Quarterly Review*, 19, 1995, pp. 2–11.

Mankiw, N.G. "The inexorable and mysterious tradeoff between inflation and unemployment." *Economic Journal*, 111, 2001, pp. C45–61.

Mishkin, F.S. "Does anticipated aggregate demand policy matter?" *American Economic Review*, 72, 1982, pp. 788–802.

Mishkin, F.S. "Is the Fisher effect for real? A reexamination of the relationship between inflation and interest rates." *Journal of Monetary Economics*, 30, 1992, pp. 195–215.

Monnet, C., and Weber, W. "Money and interest rates." *Federal Reserve Bank of Minneapolis Quarterly Review*, 25, 2001, pp. 2–13.

Mosser, P. "Changes in monetary policy effectiveness: evidence from large macroeconomic models." *Federal Reserve Board of New York Quarterly Review*, 17, 1992, pp. 36–51.

Nelson E. "Sluggish inflation and optimising models of the business cycle." *Journal of Monetary Economics*, 42, 1998, pp. 302–22.

Parkin, M. "Essays on and in the Chicago tradition." *Journal of Money, Credit and Banking*, 18, 1986, pp. 104–21.

Patinkin, D. *Money, Interest and Prices*. 2nd edn. New York: Harper & Row, 1965.

Patinkin, D. "The Chicago tradition, the quantity theory, and Friedman." *Journal of Money, Credit and Banking*, 1, 1969, pp. 46–70.

Patinkin, D. "Friedman on the quantity theory and Keynesian economics." *Journal of Political Economy*, 80, 1972, pp. 883–905.

Pesek, B.P. *Microeconomics of Money and Banking and Other Essays*. New York: Harvester Wheatsheaf, 1988.

Phelps, E.S. "Money-wage dynamics and labor market equilibrium." *Journal of Political Economy*, 76, 1968, pp. 678–711.

Pigou, A.C. "Money, a veil?" In *The Veil of money*. London: Macmillan, 1941, Ch. 4.

Prescott, E.C. "Theory ahead of business cycle measurement." *Carnegie-Rochester Conference Series on Public Policy*, 25, 1986, pp. 11–44.

Romer, D. *Advanced Macroeconomics*. New York: McGraw-Hill, 1996.

Sargent, T.J., and Wallace, N. "Rational expectations, the optimal monetary instrument, and the optimal money supply rule." *Journal of Political Economy*, 83, 1975, pp. 241–54.

Sargent, T.J., and Wallace, N. "Rational expectations and the theory of economic policy." *Journal of Monetary Economics*, 2, 1976, pp. 169–83.

Sims, C.A. "Money, income and causality." *American Economic Review*, 62, 1972, pp. 540–2.

Sims, C.A. "Interpreting the time series facts: the effects of monetary policy." *European Economic Review*, 36, 1992, pp. 975–1000.

Taylor, J.B. "How should monetary policy respond to shocks while maintaining long-run price stability? – Conceptual issues." In *Achieving Price Stability*, Federal Reserve Bank of Kansas City, 1996, pp. 181–95.

Walsh, C.E. *Monetary Theory and Policy*. 2nd edn. Cambridge, MA: MIT Press, 2003.

Wong, K. "Variability in the effects of monetary policy on economic policy." *Journal of Money, Credit, and Banking*, 32, 2000, pp. 179–98.

15 The Keynesian paradigm

The Keynesian tradition differs from the classical one in not assuming that the economy is always in equilibrium with full employment or, if out of equilibrium, tends on its own to full employment within a reasonably short time. This was the core assertion of Keynes's *The General Theory*, published in 1936, and remains at the core of all models within the Keynesian tradition. This assertion has the corollary that appropriate macroeconomic policies could improve on the functioning of the economy. Most central banks do follow this recommendation.

The Keynesian model has been evolving ever since its basis was laid in Keynes's contributions in 1936. It has gone through many versions, with different versions taking centre stage at different stages in its evolution and many coexisting simultaneously. Its latest version is the new Keynesian model.

Key concepts introduced in this chapter

- Involuntary unemployment
- Phillips curve
- Demand-deficient Keynesian model
- NeoKeynesian model
- New Keynesian model
- Notional demand and supply functions
- Effective demand and supply functions
- Sticky prices
- Staggered wage contracts
- Implicit employment contracts
- Efficiency wages
- Taylor rule
- New Keynesian Phillips curve

This chapter reviews the Keynesian ideas on short-run macroeconomics. To start, the material on the Keynesian paradigm in Chapter 1 should be reviewed. That presentation argued that the Keynesian paradigm was the study of the pathology of the macroeconomy: it focuses on the causes, implications and policy prescriptions for the deviations of the economy from its Walrasian general equilibrium position and holds that the pursuit of appropriate monetary and fiscal policies can shorten the extent and/or duration of the deviation. Since there can be

many causes of such deviations, their appropriate study requires not one unified model but many, some of which will be variations on the same theme, but there could also be models that are incompatible with one another. As such, there is no one model that can lay claim to be *the* Keynesian model. This chapter provides a small sample of the diversity of Keynesian models.

Since the classical model implies only transitory and self-correcting deviations from full employment, its emphasis tends to be on finding the long-run relationships. Since the Keynesian paradigm holds that deviations from full employment may not always be transitory and self-correcting, its focus is on finding the short-run dynamic relationships and the policies appropriate to them.[1] Keynes himself had expressed his strong disapproval of the long-run focus of classical macroeconomics in his time (and continuing at the present time) in:

> But this long run is a misleading guide to current affairs. In the long run we are all dead. Economists set themselves too easy, too useless a task if in tempestuous seasons they can only tell us that when the storm is long past the ocean is flat again.
>
> (Keynes, 1923, p. 80).

Keynesianism is a living tradition, evolving and refining its ideas over time, so that there are several versions of the Keynesian approach. The original version was Keynes's own ideas as set out in *The General Theory* (1936), followed by a number of evolving and quite diverse versions of the Keynesian framework,[2] representing a broad and often somewhat disparate set of ideas. While some of the earlier versions of Keynesianism are still part of its current mainstream, its most recent version is the new Keynesian (NK) one, which emerged only in the 1990s. Given this diversity, one needs to focus on the core themes in the Keynesian paradigm as a whole.

The common strands in the Keynesian approaches, broadly defined, are the reliance on some form of market imperfections and the angst over the performance of the economy, especially of the labor and commodity markets. Additional themes in the Keynesian literature, as in Keynes's *The General Theory*, are: the impact of changes in aggregate demand on output and employment, animal spirits (economic agents' psychology) and degree of confidence, market psychology, a consumption function that bases actual consumption on actual income through the multiplier, liquidity preference, presence of uncertainty and the possibility of default in financial markets (so that not all bonds can be assumed to be one-period riskless ones), moral hazard and contagion in financial markets, and macroeconomic instability rather than stability. While we touch on some of these topics at various places in this book, most are left to more specialized studies.

Keynesian models reject the perfectly competitive functioning of the labor market. Over time, given the complexity of this market, different Keynesian models have resorted

1 Therefore, in empirical analysis using cointegration and error-correction techniques, the concern of the classical analysis is mainly with the properties of the cointegrating vector, while that of the Keynesian paradigm is mainly with the coefficients of the error-correction equation.

2 There is considerable dispute as to whether any of the Keynesian models represents Keynes's own work. A close reading of Chapters 2 and 3 of *The General Theory* shows that they do not. It is therefore appropriate to make a distinction between the Keynesian models and Keynes's own analysis, though the former arose out of interpretations of the latter.

to different simplifying assumptions about it. At the risk of oversimplification, these were that:

(i) The nominal wage is fixed (1940s and early 1950s).
(ii) The nominal wage is variable but the supply of labor depends on the nominal and not the real wage (1950s and 1960s).
(iii) The structure of the labor market can be replaced by the Phillips curve (late 1950s and early 1960s) or the expectations-augmented Phillips curve (late 1960s).
(vi) The demand and supply of labor depend on the expected real (not nominal) wage, but the expectations on prices, needed to derive the expected real wage from the negotiated nominal wage, are subject to errors and asymmetric information between firms and workers (1970s and 1980s). Wage contracts are in nominal terms and staggered over time.
(v) The focus of the Keynesian analysis is on states other than full employment, so that the notional demand and supply functions of Walrasian and neoclassical economics are not applicable. The applicable concepts are those of effective demand and supply and the equilibrium – or "temporary equilibrium" as it is called by some writers – between them can occur at less than full employment and have different dynamic properties from those of neoclassical economics or its disequilibrium analysis. Such analysis is sometimes presented in the form of quantity-constrained models (1970s and 1980s).
(vi) In the labor market, the real wage is an efficiency real wage that can be rigid in the short run. In addition, commodity prices are sticky. Labor markets also have implicit contracts and firm-specific skills. In the commodity market, prices are sticky because of menu costs (1980s and 1990s).
(vii) The economy has forward-looking, monopolistically competitive firms that optimize intertemporally, yielding staggered, discrete adjustment of individual commodity prices. In the aggregate, this results in an NK Phillips curve relationship with the price level adjusting more slowly than in a Walrasian economy with money neutrality. Further, on monetary policy, the forward-looking optimizing central bank follows a Taylor rule for setting its interest rate (after about the mid-1990s).

The intention of this chapter is not to go through each of the various versions of the Keynesian approach but to present just three versions to show the general pattern and variety of their reasoning and implications for monetary and fiscal policies. The common theme among these versions of the Keynesian ideas is that money is not neutral in the short run and that aggregate demand management can help to improve on its performance. The models that will be used to show this are the disequilibrium or temporary equilibrium (also called the demand-deficient) model, the Phillips curve model and the NK model. Our presentation will cover the following models and ideas:

(I) The demand and supply functions are as in the neoclassical model but the labor market does not always clear, or clear fast enough, in notional terms. Further, economic agents react faster than markets to disequilibrium once it has emerged. The models used for this presentation are the effective (deficient) demand ones with involuntary unemployment.
(II) The individual equations of the labor market are replaced by the Phillips curve as a mode of encapsulating the behavior of the labor market and variations in it.
(III) There are various aspects, such as staggered nominal wage contracts, implicit contracts, efficiency wages, menu costs, etc., that affect the functioning of the labor and

commodity markets. These somewhat disparate ideas, as compared to an integrated macroeconomic model, can be grouped under the label of NeoKeynesian economics.

(IV) The final model presented is that of the integrated NK approach, encompassing an NK price adjustment process (the NK Phillips curve) and a Taylor rule to set the economy's real interest rate. This approach differs from the preceding ones by using a stochastic intertemporal general equilibrium model with monopolistically competitive firms to derive the price adjustment equations of the macroeconomic model.

Keynesian models omitted from our presentation

In particular, the models that are within the Keynesian tradition but have been *omitted* from this chapter are:

(a) Those that assume that the supply of labor depends in some way on the nominal wage and not merely on the real one. The extreme form of this hypothesis, that the nominal wage is rigid downwards, is also not presented. The latter was popular in the 1940s and 1950s and the former was common in the 1960s. Neither is now in vogue among Keynesian economists.

(b) The labor market clears but the supply of labor depends on the expected real wage, and price expectations are subject to imperfect and asymmetric information. Since such a model involves uncertainty and expectations, its presentation is to be found in the preceding chapter's analysis of the expectations-augmented Phillips curve (EAPC).

(c) Many Keynesians in the 1950s and the 1960s argued that inflation was due to the market power of monopolistic firms and/or labor unions. In some of these (cost-push) models, unions pushed for higher wages, firms followed a full-cost pricing policy and the central bank accommodated the money supply to the price level so as not to raise the unemployment rate.

On the product, money and bond markets, most Keynesians up to the mid-1990s seemed willing to accept the assumption that the central bank exogenously provides the money supply to the economy, as well as the IS–LM model's Keynesian–neoclassical synthesis of aggregate demand, which had evolved by the 1960s. The more recent (post-mid-1990s) new Keynesian models discard the LM curve, assuming that the central bank exogenously sets the interest rate or follows a Taylor rule on it, and combine it with an IS equation. The relevant IS–LM and IS–IRT models of aggregate demand in the open economy were presented in Chapter 13. This chapter takes their analysis of aggregate demand as given.

A caution on the categorization of Keynesian models

The following cautionary notes at this stage might provide some general guidance in comparing the neoclassical and Keynesian paradigms.

1 Keynesians in general accept the specification and conclusions of the classical paradigm on the long-run equilibrium of the economy. That is, in long-run equilibrium, the economy will have full employment and money neutrality. However, the Keynesian models differ from the classical paradigm in their conclusions about the short-run and short-term functioning of the economy. In particular, they focus on those deviations from long-run equilibrium that are not necessarily transient and self-correcting.

2 It is often contended that the distinguishing difference between the classical and Keynesian paradigms is that the former assumes the flexibility of nominal wages and/or prices, while the latter assumes them to be rigid. As will be shown in this chapter, the rigidity of nominal wages and/or prices is not a necessary component of some of the Keynesian versions. This is not to say that the Keynesian models cannot be based on such an assumption; it is rather to assert that they do not necessarily require such an assumption and that the qualitatively distinct Keynesian policy results can be derived with flexible prices and nominal wages provided that the condition of instantaneous market clearance is not imposed.
3 Further, the Keynesian models do not necessarily need to rely upon irrational (such as price illusion) or myopic (one-period optimization) economic behavior but can derive their distinctive conclusions under the rational demand and supply behavior of economic agents, given the conditions that come about in the relevant markets.
4 It is sometimes contended that while the Keynesian models study disequilibrium in the economy, classical models study its equilibrium properties. However, while some Keynesian models do focus on disequilibrium, others, such as the NK model, impose the requirement of general equilibrium, in which all markets are assumed to clear.
5 It is also often contended that, while the (current) models of the classical paradigm are based on micro foundations with optimal behavior by economic agents, those of the Keynesian models are not so based. This is not always so. The current crop of new Keynesian models derives its macro relationships from micro foundations with optimizing behavior.

Sections 15.1 to 15.4 present three of the main versions of the Keynesian model. Section 15.5 derives the reduced-form Keynesian relationship between output and the money supply. Section 15.6 presents the empirical assessment of the validity of the Keynesian model.

15.1 Keynesian model I: models without efficient labor markets

A market is efficient if it instantly restores equilibrium following any shift in demand or supply. Conversely, an inefficient market is one that does not have instantaneous market clearance. In the context of the labor market, note that this market is really very many diverse markets, separated by skills, location, different firms, and so on, and often with implicit long-term contracts and insider–outsider trading, etc. These factors provide plenty of scope for those who want to assume that the labor market is not efficient, while leaving others to assume that it can be taken to be efficient, or approximately so, for macroeconomic analysis.

Keynes (1936) argued that in a monetary economy the worker is generally paid a wage rate that is not guaranteed in terms of its purchasing power but is, rather, a nominal wage rate. It is the nominal wage rate that is negotiated between an employee or his union and the employer. Once a nominal wage rate has been set in an explicit or implicit contract, under normal economic circumstances neither the employee nor his union seems willing to accept a cutback in the set wage rate. However, while workers are not willing to accept a cutback in nominal wage rates, they seem more willing to tolerate a reduction in the purchasing power of their nominal wage if this is brought about by changes in the purchasing power of money.[3]

3 This is, however, a short-term phenomenon in the modern industrialized economy. Workers and their unions are nowadays sufficiently sophisticated to realize the losses in real income due to inflation and try to base the

A simple version of this ("rigid wage") model, in which the nominal wage rates are assumed to be rigid downwards, was an early version (1940s and 1950s) of the Keynesian model. This was supplanted in the 1950s and 1960s by a "nominal wage model" in which the supply of labor depended on the nominal wage rate, which was flexible and determined by the labor market.[4] However, these two models are not now seriously included among the models of the Keynesian paradigm and are omitted from this edition. Those still interested in their exposition can find the latter in Handa (2000, Ch. 15) or Patinkin (1965, Ch. 13). However, we note in passing that the assumptions of this nominal-wage model did not include the rigidity of nominal wages and prices, or of real wages, nor did it imply it. Further, this model assumed market clearance, but at the equilibrium nominal, not real, wage.

The assumption of labor supply depending on the nominal rather than the real wage implies money illusion or myopia by labor, which is empirically not valid over any extended period of time. Many Keynesians and especially post-Keynesians have argued that Keynes's *The General Theory* did not make this assumption but agreed with the neoclassical assumption that workers would base the supply of labor on the purchasing power of the nominal wage, i.e. on the real wage rate. They also argue that Keynes assumed that the modern economy, with numerous industries and firms and with decentralized wage negotiations, does not possess a mechanism that would ensure that the labor markets are normally in equilibrium at full employment with an equilibrium real wage. Hence, the claim being made under this argument is that the major distinction between the Keynesian and the neoclassical models is that, while the neoclassical model makes the assumption of labor market equilibrium as the usual state, Keynes did not do so. That is, Keynesian models of this type specify the labor market demand and supply functions as:

$$n^{\mathrm{d}} = n^{\mathrm{d}}(w) \qquad \partial n^{\mathrm{d}}/\partial w < 0 \tag{1}$$

$$n^{\mathrm{s}} = n^{\mathrm{s}}(w) \qquad \partial n^{\mathrm{s}}/\partial w > 0 \tag{2}$$

These behavioral functions are identical with those in the neoclassical model. However, on the basis of the empirical belief that the labor markets are usually not in long-run equilibrium, there is no assumption of labor market equilibrium in this type of Keynesian model. Since firms are assumed to be able to hire the amount of labor that they demand, the market clearance condition ($n = n^{\mathrm{f}} = n^{\mathrm{d}} = n^{\mathrm{s}}$) of the neoclassical model is replaced by:

$$n = n^{\mathrm{d}} \leq n^{\mathrm{s}} \tag{3}$$

Note that (3) does not assume that labor market equilibrium at full employment will never exist in the economy. This equation implicitly assumes that the modern capitalist economy does not possess sufficient mechanisms to ensure continuous full-employment equilibrium or achieve it within a reasonable period of time after a shock. Further, such an

negotiated wage on the expected rate of inflation. If the actual inflation rate is higher than the expected one, workers try to get compensation for it at the next round of wage negotiations, so that in the long run labor supply will be effectively based on the real wage rather than on nominal wages. We will return to these considerations at a later stage in this chapter.

4 The above is clearly a rather simplified picture of the wage bargain. A somewhat better picture emerges if it is assumed that workers supply labor on the basis of the expected real wage rate and the expectations are explicitly modeled. This is particularly so when nominal wage rates and prices are rising and it is difficult for workers to perceive the actual change in the real wage rate.

economy can get stuck at a level of employment below full employment, and these levels can also be equilibrium states.[5] That is, the Keynesian models allow the possibility of multiple macroeconomic equilibria, each with a different level of output and unemployment. One of these equilibria is that of full employment, so that we have to distinguish between the full-employment equilibrium and other equilibria with less than full employment.

The justification for involuntary unemployment

Designate the *long-run equilibrium* (*full-employment)* level of employment as n^{f}, and the rate of unemployment consistent with it as $u^{\mathrm{n}} \cdot u^{\mathrm{n}}$ is the natural or Walrasian (full-employment) equilibrium rate of unemployment consistent with the structure of the economy, including any wage rigidities such as specific skills, minimum wage laws, labor unions and work preferences. Also designate the *short-run equilibrium* unemployment rate as u^*, which differs from u^{n} due to errors in price expectations and the costs of adjusting prices, employment, output, etc. Separate the two components of u^* due to expectations errors, which is an element of the short-run modern classical model (see Chapter 14), and those due to adjustment costs, some of which underlie many Keynesian models (see Section 15.3), as u'^* and u''^*, respectively with $u^* = u'^* + u''^*$.

Note the identity.

$$u \equiv u^{\mathrm{n}} + (u^* - u^{\mathrm{n}}) + (u - u^*) \tag{4}$$

If there are no errors in price expectations and no adjustment costs, u^* (short-run equilibrium unemployment rate) will be identical with u^{n} (long-run equilibrium unemployment rate). Since the actual economy cannot be expected to be always and at every instant in equilibrium, or, as Keynes would have claimed, in equilibrium most of the time, $(u - u^*)$ would not always equal zero. Define $u^{\mathrm{i}} \equiv u - u^*$, where u^{i} is the deviation of unemployment from its short-run equilibrium value. With $u^{\mathrm{n}} + (u^* - u^{\mathrm{n}})$ as the short-run equilibrium unemployment rate, u^{i} will be indicative of the failure of the economy to be in either short-run or long-run equilibrium. u^{i} is usually labeled the *involuntary unemployment rate*. Its value can be positive, zero or negative, with a negative value likely to occur near the peak and a positive value near the trough of the business cycle. From the perspective of the actual economy and policy, u^{i} would be zero only if the economy, not the model, is operating in equilibrium. From the perspective of a model, u^{i} would be zero in the solution of the model if the analytical condition of short-run equilibrium is imposed in deriving the solution. Keynes and the Keynesians (other than the new Keynesians) used the former perspective and argued that the actual economy is usually not in equilibrium. Further, given their focus on deficient demand rather than on excess demand, their analysis implied positive involuntary unemployment when demand is deficient relative to the short-run equilibrium output level.

Keynesians argue, therefore, that the authorities should keep a close watch on the economy and, when there is significant involuntary unemployment due to a deficiency in aggregate demand, they should use monetary and/or fiscal policies to increase demand by an appropriate amount. If they succeed, the economy will eliminate such involuntary unemployment and perform at its full employment potential.

5 The term "equilibrium" is defined here as a state from which there is no inherent tendency to change. It specifies the values of the endogenous variables implied by the model for a specified set of values of the exogenous variables in the model.

The class of models that fit the preceding remarks are often called "*deficient-demand models*," though they allow the absence of deficient demand in the limiting case of full employment. The following subsection presents an example of a deficient-demand analysis, noting that there can also be other versions that fall into this category.

15.1.1 Keynesian deficient-demand model: quantity-constrained analysis

Keynes had argued that the economy may not always generate aggregate demand equal to the full-employment supply of commodities. The former comes from the decisions of very many households, firms and the government (and exports in the open economy) in the form of consumption, investment and government expenditures. The latter is generated by firms, which undertake production to meet expected sales. There is no a priori reason for the two to be equal. The two could, however, be brought into equality by efficient markets that adjust the price to the equilibrium level. However, Keynes argued that the markets were not efficient (i.e. following a shift in demand or supply, adjusting the price instantly to equate demand and supply), so that firms and households would make their production and consumption decisions in response to expected demand and incomes, rather than to changes in the market price. In brief, markets were sluggish in their adjustment, and firms and households reacted faster than markets. This implies that firms would base their output and investment on expected demand, so that changes in aggregate demand would change aggregate output and employment. Households, in deciding on their consumption expenditures, reacted to their expected income and job prospects and not merely to their real wage rate, so that their consumption, and therefore aggregate demand, would respond to their job prospects. Hence, a fall in aggregate demand would reduce output and employment, which would, in turn, worsen incomes and job prospects and lower consumption (Clower, 1965; Patinkin, 1965; Leijonhufvud, 1967, 1968).

Writing in the Great Depression, Keynes argued that the economy often functions below its full-employment level. While a depression has not been experienced since the 1930s by the major economies, all economies do suffer periodic recessions. Keynes's factual statement is generally taken to be correct for recessions.[6] The most plausible reason for output falling below its full-employment level is a demand deficiency originating with a fall in aggregate demand. To show this, start with the initial state of full-employment equilibrium in the economy and assume that aggregate demand falls for some reason, thereby creating deficient demand relative to the full-employment level. Following the ideas of the preceding section, assume that the labor market does not instantly clear, so that some of the workers become involuntarily unemployed because of the fall in aggregate demand. These workers will not receive the wage they would have got if they had been able to sell their labor according to their supply curve. Their lack of income forces them to reduce their demand for commodities and real balances below that specified by their Walrasian functions as derived in Chapter 3 and the neoclassical functions specified in Chapter 14. Hence, these latter functions are not "effective" – that is, operational – in the state of involuntary unemployment. They can only be effective if there is full employment in the economy and workers could sell all the labor they wanted to at the existing wage rate.

6 It is a well-established stylized fact that virtually all recessions have a fall in aggregate demand and a fall in output, as well as a fall in the inflation rate.

Demand and supply functions derived under the assumption of the simultaneous clearance of all markets are called *notional functions*.[7] They are the ones derived in Walrasian analysis and used in neoclassical models. Since involuntary unemployment means that at least one of the markets does not clear, the use of notional functions to analyze the existence or non-existence of involuntary unemployment begs the question and is inappropriate. The more appropriate analysis would be to posit that the real-world – that is, actual – demand and supply functions are approximated by effective functions, of which the limiting case is notional functions.

Effective functions for any market that take account of the non-clearance of other markets are also called *Clower* or *quantity-constrained functions* and the macroeconomic analysis based on them is similarly called quantity-constrained analysis. Such analysis clearly belongs among the Keynesian stable of macroeconomic models and became popular during the 1970s and 1980s.

Quantity-constrained analysis can encompass the possibility that any or all of the four markets of the macroeconomic models need not clear instantaneously. However, such Keynesian analysis focuses on non-clearance in the commodities[8] and labor markets. In particular, the main initial impulse of such models is a fall in the aggregate demand for commodities due to a fall in investment, in consumption, in government expenditures or exports, or in money supply.

Dynamic analysis following a fall in the aggregate demand for commodities

While the reader is referred to an appropriate macroeconomics book for a proper treatment of effective demand models, we pursue here Patinkin's (1965, Ch. 13)[9] application of this analysis to the market for labor in order to derive the role of monetary policy in such models. Assume that the demand and supply functions for labor are the neoclassical notional ones, as shown in Figure 15.1b, and that initially the economy is at full employment n^f in Figure 15.1b and full-employment output y^f in Figure 15.1a. Now assume that a shock reduces aggregate demand to $y^d{}_0$ so that a demand deficiency emerges in the economy such that the firms are not able to sell the full-employment output y^f at the existing price level.[10] The actual aggregate demand $y^d{}_0$ can be supplied by the employment of $n^d{}_0$ workers. In Figure 15.1b, the marginal product of labor for $n^d{}_0$ workers is MPL_0, which is above the full-employment wage of w^f. However, if firms were to employ more than $n^d{}_0$ workers, they would not be able to sell

7 See Chapter 18 on Walras's Law for additional exposition on these functions.

8 Such inequality of demand and supply occurs in notional terms, whereas for the commodity market the equality of the actual or effective demand and supply would still occur and determine prices. For the labor market, such non-clearance means that the notional supply of labor exceeds the notional demand.

9 Patinkin was one of the most distinguished contributors to the neoclassical approach, even though his contribution on the following analysis of deficient demand was in the spirit of Keynesian economics.

10 Neoclassical reasoning at this point would be that the aggregate price level would fall, creating a real balance effect which shifts the LM curve to the right, thereby increasing incomes and inducing an increase in consumption spending, so that the fall in aggregate demand would be reversed. Keynesians claim that this process does not work or is too slow, for many reasons: uncertainty of whether the fall in demand is transitory or longer lasting; firms may find it optimal not to change their price lists immediately; the real balance effect is quite ineffective and extremely slow in increasing demand; etc. Hence, Keynesians tend to assume that, for analytical realism, it is better to assume constant prices rather than falling prices, so that the response of firms to the fall in demand is to adjust the quantity produced rather than to reduce prices. The following analysis follows this Keynesian procedure.

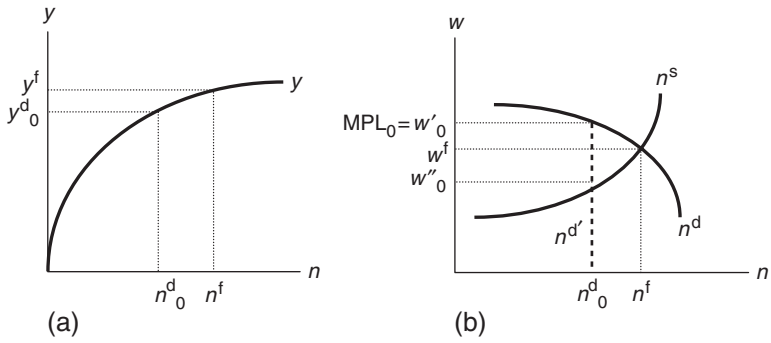


Figure 15.1

the extra output so that their *marginal revenue product* would be zero. Hence, if aggregate demand fell to y^d_0, firms would cut employment to only n^d_0 workers.[11]

Extending this argument to Figure 15.1b, the employed n^d_0 workers can be paid nominal wage rates which can change, as can the price level, with the resultant real wage being anywhere within the range w'_0 and w''_0, without a change in the firms' employment of n^d_0 workers, so that real wage rates could drift up or down from the initial equilibrium level of w^f.[12] Hence, the decrease in employment (from n^f to n^d_0) can be accompanied by either an increase or a decrease in the real wage rate of the employed workers. Wages may therefore follow a pro-cyclical or counter-cyclical pattern: some recessions and some parts of a given recession could show wages falling while others show them to be rising. If wages rise, it could be claimed that the rise in wages is the cause of falling employment, when this rise is itself only an effect while the true cause was the initial fall in aggregate demand.

The above effects are only partial or initial ones. Since the unemployed workers do not receive any income, they cut back on their consumption demand. Further, if wages were cut below w^f, the lower incomes of the employed would also lead to a reduction in consumption.[13] The consequent fall in aggregate demand further reduces the effective demand for labor[14] and acerbates the recessionary effects derived in the preceding paragraph.

The essential components of the preceding analysis are:

(a) The economy's industrial and market structures do not lead to instant market clearance in all markets, so that rational firms cannot assume this to be so, and must react in their

11 As against this dynamic reaction, neoclassical economics claims that firms will cut prices, not output, in response to a fall in aggregate demand; or that markets will act fast enough to deliver the lower prices for all commodities sufficiently to allow firms to sell all they want before individual firms react by cutting their production. Intuitive knowledge of the economy tends to favor the Keynesian response pattern.

12 Real wages will rise if the price level falls faster than nominal wages, they will fall if the price level falls more slowly than nominal wages, and will stay constant if both prices and wages fall in the same proportion.

13 The fall in employment usually increases, among the employed workers, the subjective risk of staying employed, so that such workers also tend to cut back on consumption in order to increase precautionary saving to provide for the eventuality of becoming unemployed.

14 The new demand curve is not the neoclassical notional one n^d but an effective one, and one cannot proceed with analysis using n^d.

production and labor demand strategy to emerging demand conditions for their products as they see fit. They react faster than markets to changes in demand for products and labor.

(b) There is no coordinating mechanism or coordinating agency (such as the Walrasian auctioneer with costless and open recontracting while adjustments to a new equilibrium are going on) in the economy which will work out the equilibrium wages, employment, output and prices instantly following shifts in demand and supply. Markets are competitive and tâtonnement towards equilibrium values may occur, but takes time, during which firms and workers make decisions on employment.

(c) There is no nominal wage or price rigidity. Prices and wages would be market determined if the markets were efficient or at least adjusted faster than economic agents in response to shocks.

(d) Firms and households are rational and react to the conditions that they face in the absence of perfect markets.

The lack of perfect market structures ensuring instantaneous adjustment to equilibrium or the lack of a coordinating mechanism or agency – with firms and workers not waiting out this process but responding faster than the sluggish markets to a fall in demand is the critical reason in effective demand models why an economy in which a demand deficiency has emerged need not rapidly move to the full-employment level. Critical to the dynamic responses of economic agents are their rationally expected values (of aggregate demand, prices and employment) that apply in the period (day, month, quarter, etc.) ahead. Their estimates of these expectations are reflected in some way by economic/business analysts' indices of business and consumer confidence.

It seems reasonable to posit that a very mild fall in aggregate demand in an overall time path of full employment can leave both business and consumer confidence intact and not produce a reduction in output and employment, whereas a more significant one or/and over a longer period, especially when backed by past experienced recessions, would produce a loss in such confidence and take the economy onto a dynamic path to involuntary unemployment levels. The irony of this remark is that the deliberate pursuit of aggressive Keynesian policies to maintain aggregate demand at the full-employment level leads to a dynamic response by consumers and firms which maintains full employment; but a policy of leaving the economy alone, as the classical economists recommend, brings into play dynamic responses which take the economy away from full employment. In support of this contention, the Western economies experienced very shallow and short recessions during the Keynesian period of the 1950s and 1960s, whereas the recessions lengthened and worsened with the resurgence of classical economics in the 1980s and 1990s.

Note that new Keynesian models of the post-1990 variety replace the assumption that economic agents react faster than sluggish markets by the assumption that firms are price setters because they are in monopolistic competition.

Optimal monetary policy in the Keynesian demand-deficient economy

To derive the optimal role and impact of monetary and fiscal policies in such a context, assume that the economy is now at $n^{d}{}_{0}$ of employment and $y^{d}{}_{0}$ of output in Figures 15.1a and 15.1b, and suppose the monetary authorities pursue an expansionary monetary policy. This increases aggregate demand in the economy, thereby shifting the n^{d} curve from $n^{d'}$ towards n^{f} and increasing output beyond $y^{d}{}_{0}$. Since output increases in response to the increase in

aggregate demand, prices may or may not increase. The increase in prices will depend on how deficient the earlier demand was and how large was the earlier excess capacity in the economy, but, in any case, prices will not rise in proportion to the increase in the money supply.[15] The expansionary monetary policy would have succeeded in increasing output in the economy. But once the economy has reached y^f – that is, there no longer exists any demand deficiency further expansions in the money supply will not increase output but merely cause a proportionate increase in the price level. This limiting case of demand-deficient analysis is, of course, the neoclassical full-employment case, in accord with the Keynesians' claim that their analysis is the more general one and encompasses the neoclassical full-employment and classical cases as a limiting case.

These arguments imply that there is no straightforward or linear relationship between increases in the money supply and real output, or between the rate of inflation and the level of unemployment. These relationships depend upon the state of the economy and the extent of the monetary expansion. Further, the transmission of the impact of money supply increases on output does not always require or go through price level increases.

The economy's responses to excess demand

We have so far considered the dynamics of a decline in demand from a full-employment level. But suppose aggregate demand increases when there already exists full employment and the economy starts with the natural rate of employment. The dynamic response patterns of firms and workers should be basically consistent, though in opposite directions, in the two situations since these would be based on their rational response behavior patterns, but the economy's constraints would be quite different between these cases. On the former, the individual firm, seeing an increase in the demand for its product, would tend to increase production through increases in employment, overtime worked, increased effort of employees and increases in efficiency. Each of these is feasible in the short run, with the increase in employment, beyond an initial full-employment level, occurring through increased working hours of part-time workers, through students delaying resumption of studies and through increases in the overtime put in by employed workers, etc.[16] While such increases in employment and output above their full-employment levels can and do occur, as long drawn-out booms indicate, their scope is constrained by the economy's short-run flexibility for these variables. Hence, while the increases in aggregate demand can and do increase output and employment in the short run beyond their full employment levels – and the economy can go below its natural rate of unemployment – the increase in prices is more likely, and faster, than the fall in prices in response to a decline in demand from the full-employment level.

Most central banks now believe that changes in aggregate demand do produce changes in the economy's output and that the economy sometimes operates below full employment and sometimes above it. Evidence of these is provided by the current popularity of the Taylor rule, which tries to limit the gap between actual output and full-employment output, as well as the deviation of inflation from a target level, by the impact of monetary policy on aggregate demand.

15 The real wage may rise or fall, depending upon where it was earlier in relation to w^f.

16 However, they cannot be sustained over time (e.g. workers get tired of putting in undesired overtime) and such increases in employment cannot be assumed for the long run.

The eclipse of the deficient-demand analysis

The above deficient-demand analysis was bypassed by new developments, especially the new Keynesian modeling, in the Keynesian paradigm during the last quarter of the twentieth century. One reason for this was that its theoretical development had reached a dead end. Further, it was not intellectually challenging in the context of precise mathematical modeling, especially of macroeconomic dynamics, which had come to dominate macroeconomics. An additional reason was the adoption by the new Keynesians of the intertemporal general equilibrium methodology under imperfect competition, which, by the nature of its assumption of general equilibrium, excluded the dynamic effects of deficient demand. However, this intellectual shunting aside of deficient-demand analysis does not necessarily affect its validity, so that it remains a component analysis of the Keynesian paradigm.

15.2 Keynesian model II: Phillips curve analysis

Phillips curve

In 1958, A.W. Phillips, on the basis of statistical observations for the UK proposed a negative relationship between the rate of nominal wage growth and the rate of unemployment. This was subsequently extended to show a negative relationship between the rate of inflation and the rate of unemployment, with the name "Phillips curve" being attached to both of these relationships. During the late 1950s and the 1960s, Keynesian economics came to embrace this curve, incorporating it in preference to a structural specification of the labor market, as in equations (1) to (3).

Phillips (1958) had plotted the rate of change of nominal wages against the rate of unemployment for the UK over several periods from 1861 to 1957, and found that the data showed a downward-sloping curve. That is, the plotted relationship was of the form:

$$\overset{\circ}{W} = f(u) \tag{5}$$

where $\overset{\circ}{W}$ is the rate of increase of the nominal wage rate and $f' < 0$. One explanation for (5) is that unemployment represents the degree of labor market tightness, so that the higher the level of unemployment, the smaller will be the increase in the nominal wage.

Equation (5) soon evolved into its inverse and then into a relationship of the form:

$$u = g(\pi) \tag{6}$$

where $g' < 0$. This relationship is drawn as the PC curve in Figures 15.2a and 15.2b.

The transition from (5) to (6) comes from the link between the nominal wage rate and inflation: nominal wages represent the main element of the cost of production, so that an increase in nominal wages will induce firms to increase their prices; alternatively, an increase in prices causes labor to ask for compensation in the form of wage increases. Hence, there is a positive relationship between $\overset{\circ}{W}$ and π, which, when substituted into (5), yields (6).

The estimated forms of the Phillips curves proved to be convex to the origin. To explain this curvature, it was argued that the response of nominal wages to excess demand was non-linear, and that decreases in unemployment caused successively greater increases in nominal wages. Further, a decrease in employment induces a smaller decline in wages than the increase in wages brought out by a corresponding increase in job vacancies, so that increases

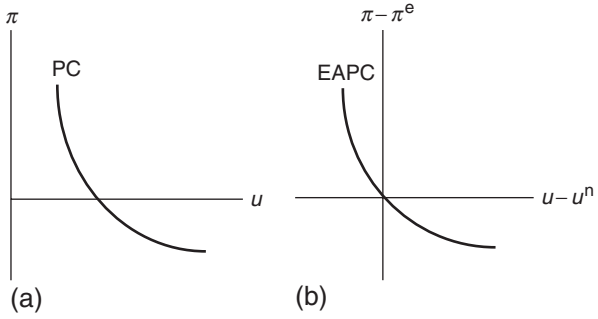


Figure 15.2

in employment in some industries with corresponding decreases in others will bring about a net increase in the average nominal wage rate. Hence, both the level of unemployment and its variance among industries together determined the Phillips curve relationship.

For the pre-1970s data, numerous studies for many countries, including Canada and the USA, seemed to confirm the validity of the Phillips curve. Even though the relationship seemed to differ between periods and countries, the general form of the relationship seemed to be valid for the 1950s and 1960s and became a mainstay of many Keynesian models, replacing the labor market structural relationships, during the 1960s and 1970s. As a consequence, many Keynesian economists in the 1960s and early 1970s assumed the Phillips curve to be stable and recommended its use as a trade off between inflation and unemployment by the monetary and fiscal authorities, calling on them to use their policies to change aggregate demand to achieve the inflation rate specified by the Phillips curve as a concomitant of the desired rate of unemployment in the economy. In this sense, while (6) was a constraint on policy choices, the Keynesians interpreted it as giving the authorities control over the unemployment rate in the economy, with the accompanying rate of inflation being an undesirable cost of the chosen unemployment rate. The Phillips curve provided the economic tool to support the economic philosophy of the Keynesians in the 1960s and 1970s that the monetary and fiscal authorities should try to achieve better levels of output than the economy would generate on its own, even though doing so would mean higher inflation rates.[17]

The expansionary monetary and fiscal policies resulting from central banks' attempt to take advantage of the Phillips curve tradeoff did lead to rising rates of inflation.[18] Once the inflation reached unacceptable levels, with a momentum towards further increases, the central banks would resort to monetary stringency to fight it. However, by this stage, the public had come to expect higher inflation and the inflation-fighting monetary contractions resulted mostly in increases in unemployment, reaction to which could force the central bank to again

17 There was also at that time considerable skepticism among the Keynesians on whether the central bank could control inflation, since it was often attributed to cost-push forces, or because it could only control the money supply, which was only a small part of liquidity in the economy. The latter point was discussed in Chapter 2 under the topic of the Radcliffe report.

18 Boschen and Weise (2003) explore the origins of 73 inflation episodes in OECD countries from the 1960s to the 1980s. They report that inflation often started because, out of a belief in the short-run Phillips curve, central banks sought the short-term benefits of higher growth without considering the costs of disinflation later.

pursue a monetary expansion, thereby producing a "stop–go" policy pattern. The role of expectations proved vital to the impact of changes in inflation on output and unemployment, and led to the modification of the Phillips curve to the expectations-augmented curve.

The expectations-augmented Phillips curve[19]

The Phillips curve was challenged by Milton Friedman and the monetarists in the 1960s and 1970s. They argued that if inflation were perfectly anticipated, the labor contracts would reflect it so that the nominal wage would increase by the expected rate of inflation. Consequently, the expected inflation rate would not affect the real wage rate, employment or output. Hence, at the expected inflation rate, the unemployment rate would be the natural rate so that only the unanticipated rate of inflation would cause deviations in the actual from the natural rate by reducing the real cost of labor and other inputs. That is, according to Friedman, the proper relationship between u and π is not (6) but has the form:

$$(u^* - u^{\mathrm{n}}) = f(\pi - \pi^{\mathrm{e}}) \tag{7}$$

with $f' < 0$ and $f(0) = 0$. u^{n} is the long-run equilibrium unemployment rate, while u^* is the short-run equilibrium unemployment rate in the presence of errors in inflationary expectations. In the limiting case of expectational equilibrium, an aspect of the "long run," $\pi = \pi^{\mathrm{e}}$ and $u^* = u^{\mathrm{n}}$. (7) is the expectations-augmented Phillips curve, as shown as the curve EAPC in Figure 15.2b. Its analysis and policy implications were presented in Chapter 14.

Empirical research and the widespread experience of stagflation in the mid and late 1970s in the industrialized economies seemed to show that (6) was unstable in a period of accelerating inflation. Further, (7) seemed to perform much better in such periods, especially at high and accelerating rates of inflation. In response to this analysis and evidence, Keynesian models after the 1970s tended to drop the Phillips curve as a primitive element of their models, though most such models can generate some form of it as an implication. Its most frequent resurgence is in the form of the new Keynesian Phillips curve, which is rather more in the nature of firms' output–price adjustment equation than a true Phillips curve based on labor market behavior.

The relationship between the Phillips curve and the expectations-augmented Phillips curve

To look at the relationship between the actual Phillips curve and the expectations-augmented Phillips curve, start with the identity:

$$u \equiv u^{\mathrm{n}} + (u^* - u^{\mathrm{n}}) + (u - u^*)$$

where u is the actual unemployment rate, u^* is the short-run unemployment rate in the presence of errors in price expectations and u^{n} is the long-run equilibrium unemployment rate. From the Friedman and Lucas analyses in Chapter 14, we have:

$$u^* - u^{\mathrm{n}} = f(\pi - \pi^{\mathrm{e}})$$

19 The formal derivation of this curve is presented in Chapter 14, which discusses uncertainty and expectations.

Therefore,

$$u = u^{n} + f(\pi - \pi^{e}) + (u - u^{*}) \tag{8}$$

The Phillips curve focuses on the determinants of $(u - u^{n})$. There could be many determinants of this difference. One of these relies on errors in price or inflationary expectations. Keynesian models provide several other determinants, with only one of them being the demand-deficient analysis. New Keynesian models, presented later in this chapter, provide another form of the Phillips curve, which arises from monopolistic competitive behavior in commodity markets and price adjustment costs.

To emphasize the rarely recognized point about (8), the expectations-augmented Phillips curve is a component of the Phillips curve analysis, but may not even be the most significant part empirically. However, in view of this relationship between the curves, the Phillips curve does shift if there is a change in expectations of inflation, but this is only a part of the story.

The Phillips curve as specified by (6) or (8) is also different from the new Keynesian Phillips curve discussed later in this chapter.

15.3 Components of neoKeynesian economics

We group under this heading the economic theories developed since the 1970s to provide a foundation for the Keynesian tenets that involuntary unemployment can exist in the economy and that changes in aggregate demand in the economy can affect output, at least in the short run. These theories have in many ways supplanted the traditional Keynesian (nominal wage, demand-deficient and the original Phillips curve) models presented in the earlier sections of this chapter. We will present three of these theories. They are the efficiency wage theory leading to real wage rigidity, a theory of rigid or sticky prices based on sluggish price adjustments, and a theory of implicit contracts leading to labor hoarding. These are used in some combination to derive the neoKeynesian conclusions that monetary policy, through aggregate demand, affects output and is not neutral in the short run, though it is neutral in the long run.

15.3.1 Efficiency wage theory

While the neoclassical and Keynesian theories presented so far in this book have taken the effort put in by each worker on the job to be constant, the efficiency wage model proposed by Akerlof (Yellen, 1984; Shapiro and Stiglitz, 1984) assumes that this effort is a function of the worker's wage. It can also be a function of other variables such as the unemployment rate and the unemployment benefits that also affect the opportunity cost of the job. In order to accommodate the concept of a variable effort on the part of workers, the firm's production function is modified from the usual one of:

$$y = f(n, \underline{K}) \qquad f_n > 0,\ f_{nn} < 0$$

to:

$$y = f(e(w)n, \underline{K}) \qquad f_e,\ f_n > 0,\ f_{ee},\ f_{nn} < 0,\ e_w > 0, e_{ww} < 0 \tag{9}$$

where e designates "effort," taking this to be a measurable variable. (9) keeps the capital stock constant for short-run analysis. Effort e is a function of the real wage rate w. Paying the

workers more than the market clearing wage tends to increase their productivity by reducing shirking by workers, reducing turnover of workers, improving the average quality of job applicants and improving morale in the firm.

Focusing only on the shirking element, most jobs do not rigidly force the workers to work at a pre-set pace with a pre-specified productivity but allow them some leeway in their level of performance. Workers could, therefore, "shirk" on the job, thereby lowering their productivity or requiring the firm to prevent shirking through performance monitoring by "inspectors," which imposes additional costs on the firm. If the worker is paid the market wage only, and is fired if caught shirking, he could obtain another job at the same wage and therefore would only lose by the extent of his search costs. But if he is paid a wage higher than the market wage, he has an incentive not to shirk and thereby lose a job with a better wage than he can get if he shirked and was fired for doing so. The absence of shirking, in turn, increases the worker's effort and productivity. The firm therefore has an incentive to pay its workers more than the market clearing wage. For optimization, the wage paid will be that which yields the lowest labor cost per efficiency unit – that is, with labor measured in terms of its efficiency. Designate this optimal wage, known as the efficiency wage, as w^*. The profit-maximizing firm will then employ labor up to the point at which its marginal product equals its real wage, that is, by:

$$\partial y/\partial n = e(w^*)f'(e(w^*)n, \underline{K}) = w^* \qquad f'(e(w^*)n, \underline{K}) = \partial f/\partial(e(w^*)n) > 0 \tag{10}$$

In equilibrium, all firms would pay the real wage w^*, assuming it to be greater than labor's reservation wage.

The efficiency wage models assume the neoclassical labor demand and supply functions, with both demand and supply being functions of the real wage, so that the labor market is represented by Figure 15.3. At the wage w^*, higher than the market clearing wage, $n^{d}{}_0 < n^{f} < n^{s}{}_0$, so that employment at n^* is less than full employment and there exists involuntary unemployment equal to $(n^{s} - n^{d})$. These unemployed workers are willing to accept the wage w^* or w^{f} or even somewhat lower wages but their competition for jobs will not reduce the market wage, since such a lower wage will be below the efficiency wage w^* and reduce the productivity of firms' existing employees and their profits. Consequently, the labor market will maintain involuntary unemployment equal to $(n^{s} - n^{d})$ in the long run. Since such involuntary unemployment is a long-run feature of the labor market, it can be called the long-run involuntary unemployment. Because of its long-run nature,

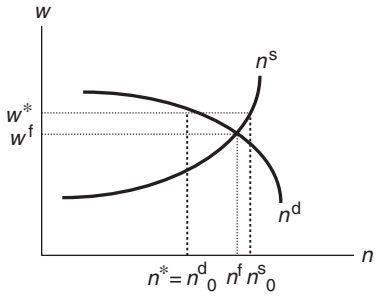


Figure 15.3

some economists propose its incorporation into the notion of structural unemployment and redefine the natural rate to encompass it. However, since the determinants of structural unemployment and of such long-run involuntary unemployment are quite different, we prefer to keep them as separate concepts. Nor do we merge the latter into the classical concept of the natural rate of unemployment. In the efficiency wage context, the long-run rate of unemployment will, therefore, be the classical natural rate plus efficiency-based long-run involuntary unemployment.

To explain the effects of a fall in demand on employment, rewrite (10) as:

$$p_i e(w^*) f'(e(w^*)n, \underline{K}) = Pw^* = W^* \tag{11}$$

where p_i is the price of the firm's product, P is the price level and W^* is the nominal wage equal to Pw^*, with w^* still the efficiency real wage. (11) can be rewritten as:

$$e(w^*) f'(e(w^*)n, \underline{K}) = (P/p_i)w^* \tag{12}$$

Now assume that there is a decline in the demand for the firm's product such that its relative price (p_i/P) falls. This will not change the firm's efficiency real wage w^*, but its demand curve for labor will shift down and its employment will fall. Conversely, an increase in the relative price of the firm's product will leave the efficiency real wage unchanged but increase its demand for labor and employment.

Now suppose that all product prices increase proportionately, so that we can use either (10) or (12). These equations imply that a change in the price level will not bring about a change in the efficiency wage w^* or the equilibrium values of employment n^* or involuntary unemployment u^{i*} in the economy. That is, this version of the efficiency wage theory does not imply that a change in aggregate demand – due to monetary or fiscal policy or other exogenous changes – will change aggregate employment and output. For this, we need to bring in the neoKeynesian theory of price stickiness, presented in the next subsection.

The efficiency wage theory implies long-run involuntary unemployment in the economy.[20] It can also be used to buttress the Keynesian claim that, following a fall in demand, the economy could move to an equilibrium with a still higher level of unemployment. For this, it is preferable to modify the effort function to a more realistic one as $e = e(w, u)$, where $\partial e/\partial w \geq 0$ and $\partial e/\partial u \geq 0$, so that as unemployment rises, the firm can lower the efficiency wage necessary to get the maximum effort from the workers. In this case, a tradeoff will exist between the optimal efficiency wage and unemployment in the economy. A given fall in aggregate demand will, then, imply an equilibrium with a higher level of unemployment and a lower efficiency wage.

15.3.2 Costs of adjusting employment: implicit contracts and labor hoarding

The neoKeynesians also argue that it is optimal for firms and workers to enter into long-term implicit and explicit employment contracts. Such contracts are optimal for the firm because of the cost of hiring and training workers and the firm-specificity of skills acquired through training and learning on the job, so that the productivity of such a skilled worker will be

20 It also explains the existence of dual markets, wage distributions among workers with identical skills and certain types of wage and job discrimination.

greater than of new hires. The worker also benefits from this higher productivity through higher wages in his existing firm than if he were to quit and join other firms. This mutual benefit from continued employment implies that the firm will try to retain its skilled workers if it can do so through a period of reduced demand for its output, rather than laying them off immediately. The firm therefore finds it optimal to lay off fewer workers than is justified by the fall in demand, leading to a form of labor hoarding during recessions (Okun, 1981). Such hoarded labor works less hard during recessions because there is less work to do, or is often diverted to such tasks as maintenance. If a worker is laid off, he also has an incentive to wait to be recalled by his old employer rather than immediately accept a job with another firm in which his productivity and wage will be lower. Hence, reductions in aggregate demand in the short run partly lead to labor hoarding, with a consequent fall in average productivity, and partly to an increase in unemployment, with some of the laid-off workers being put on recall and voluntarily waiting to be recalled rather than actively searching for jobs.

Conversely, the implicit agreement between firms and workers also means that workers accommodate increases in demand for the firm's output with increased effort, even in the absence of wage increases. Hence, output fluctuates more than employment over the business cycle, and the fact that the economy is in its long-run full-employment state is not a barrier to short-run increases in output.

15.3.3 Price stickiness [21]

NeoKeynesian theory assumes that while some goods in the economy are homogeneous and are traded in perfectly competitive markets, most goods, especially at the retail level, are differentiated by firms in some characteristic or other. Such differentiation is often in the form of differences in color, packaging, location, associated services, or just established brand loyalty. Such differentiation in practice is usually not enough to create a monopoly for the firm but enables it to function in a monopolistically competitive manner. Profit maximization by a monopolistically competitive firm implies that it is not a price taker, as are firms under perfect competition, but a price setter with a downward-sloping demand curve for its product. Consequently, increases in the price it sets do not reduce its sales to zero, nor do reductions in it allow it to capture the whole market for its industry. As a price setter, the firm sets the profit-maximizing price and supplies the output demanded at this price.

Changing the set price imposes a variety of costs, collectively known as *menu* costs. Examples of these are: reprinting price lists and catalogues, informing customers, re-marking the merchandise, etc. These costs, though often relatively small as a percentage of the price of the firm's product, can still be greater than the gain in revenue from a small price change. Further, even if there is a net gain from changing the price following an increase in demand, it may not be enough to persuade the firm to immediately raise its price, since the inconvenience and costs to the firm's customers of frequent price changes are likely to be resented. Consequently, the firm may not find it optimal to respond to demand changes with price changes unless the demand changes imply large enough price changes. Over time, as demand increases occur, the optimal price change becomes large enough for the firm to be willing to incur the menu costs and change the actual price of its product.

These arguments imply that a monopolistically competitive firm will change its price infrequently, but will respond to intervening changes in demand by changing its output at the

21 For an exposition of this approach, see Ball *et al.* (1988).

existing price. In the long run, the price will adjust to demand and, even in the short run, if the demand increase is large enough, the price adjustment will occur. In the economy as a whole, an increase in aggregate demand will cause some sectors and firms, especially those with more competitive markets, to adjust their prices faster, while others will not immediately do so but will respond to demand changes with supply changes. Consequently, an increase in the aggregate demand will be partly met with an increase in prices and partly by an increase in output.

The increase in aggregate output to meet an increase in aggregate demand requires an increase in employment. Even if the economy was initially in its long-run state, the efficiency wage theory, unlike the classical theory, implies that there would exist, in this long-run state, the involuntary unemployment of workers and that these workers are willing to accept jobs at the existing real wage. Hence, the increase in aggregate demand will be accommodated through an increase in employment and output, without necessarily a change in real wages.

Conversely, a decrease in the demand for the products of the firm in monopolistic competition need not immediately cause it to lower its price unless the implied optimal price reduction was sufficiently large. Again, taking the economy to be a mix of firms in perfect and monopolistic competition, decreases in aggregate demand would partly result in a fall in the price level and partly in a reduction in the output supplied. The latter will cause firms to reduce their employment. However, as the efficiency wage theory argued, this fall in employment need not lead to a competitive reduction in the real wage.

Under the sticky price hypothesis, macroeconomic fluctuations, sometimes large ones, can arise from even small menu costs. Suppose that, because of a reduction in the money supply, aggregate demand falls and there is a corresponding fall in each firm's demand. If a given firm lowers its price, it moves along the new demand curve and not the old one. Its gain in revenue from this movement is relatively small and, because of menu costs, a lower price may not increase its profits, so that it does not reduce its price. Instead, it reduces its output to meet the new, lower, demand at the pre-existing price. With each firm behaving in this manner, aggregate output will fall. However, if all firms reduce their prices simultaneously, the price level will fall, real money balances in the economy will rise and the economy's and the firm's demand curves will shift back to their original position, so that there will be no drop in the firm's or the economy's output.

Sticky prices provide a justification based on menu costs for an upward-sloping short-run aggregate supply curve. This justification is different from that for the Phillips curve or its expectations-augmented version, and is also different from that for the Lucas supply rule. Its extreme version is the simplification that no firms would change prices in response to demand changes, so that the aggregate price level can be taken as constant. In this version, shown in Figure 15.4a, the price level is constant at its initial level P_0 given by the point a. The LAS curve shows the aggregate supply as y^f. Prices are sticky at P_0, so that we can specify a short-run aggregate supply curve SAS which is horizontal at P_0. An increase in aggregate demand from AD_0 to AD_1 leads to the supply of output y_1 at the sticky price P_0. Conversely, a decrease in the aggregate demand from AD_0 to AD_2 leads to the supply of output y_2, but again without an accompanying change from the sticky price level P_0. Transient and small changes in aggregate demand are therefore accommodated by the change in output, with an accompanying employment change. Cumulative changes in the same direction – or large aggregate demand changes – will, however, cause increases in prices, so that the long-run response to such changes is taken to be along the LAS curve.

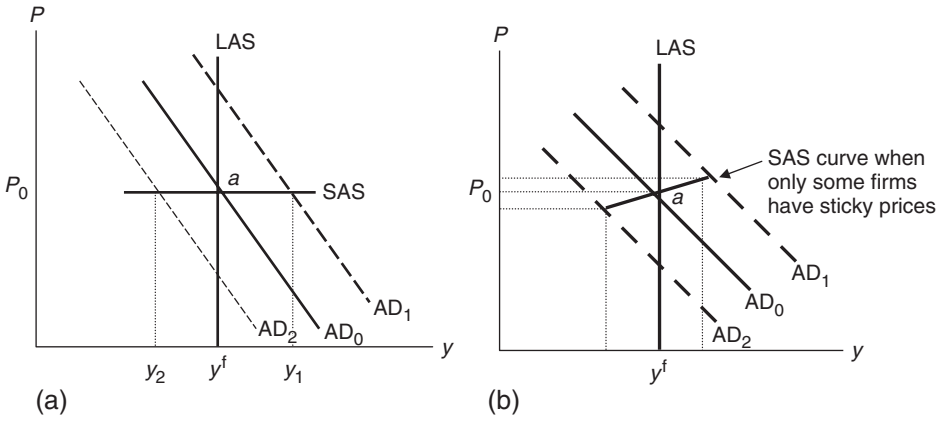


Figure 15.4

An economy with a mix of perfectly competitive (without sticky prices) and monopolistically competitive industries will have an SAS curve that is upward sloping rather than horizontal at the initial price level. Further, not all firms with sticky prices will experience sticky prices at the same time, so that the revision of prices will be staggered over time.[22] These arguments imply that the aggregate supply curve will not be horizontal but will have a positive slope, with this slope being less than if none of the firms experienced sticky prices. The nature of this curve is shown in Figure 15.4b. It assumes that outside a certain range, defined by AD_1 and AD_2, of the change in aggregate demand, it becomes profitable for all firms to change their prices. Within this range, the short-run aggregate supply curve has a positive slope with some prices remaining unchanged.

The term "menu costs" suggests small costs of changing individual prices. However, as shown above, these small costs can have "large" implications in terms of the impact of monetary policy on output and the departure from its neutrality. This result is sometimes used to make the assertion: "small nominal rigidities can have large real effects."

An implication of this menu cost approach to price stickiness is that since the firm's decision to change its price, given the menu costs, rests on whether the price change will increase or reduce its profits, it will adjust its price faster the greater the change in demand. The quicker prices adjust, the smaller will be the effect of aggregate demand increases on real output. Hence, larger changes in demand are likely to produce smaller real effects and, beyond a certain point, no increase in output. Sticky prices also reduce the real impact of demand increases in an inflationary environment. In a high inflationary environment, since prices are rising anyway, sticky price firms will find it profitable to adjust their prices more rapidly than in a zero or low inflation environment. This will reduce the increase in output.

22 The staggered nature of this process can be captured in a Calvo adjustment process (Calvo, 1983) under which the representative firm decides on the probability $(1-\theta)$ that it will adjust its price this period and the probability θ that it will not do so. Assuming these probabilities to be independent of the time when it last adjusted its price, the average time for which the representative firm's price remains fixed is $1/(1-\theta)$.

Importance of staggered price and wage adjustments

NeoKeynesian theories and the new Keynesian model (see below) assert that price (and nominal wage) adjustments are staggered over time rather than occurring simultaneously in a context of price-setting firms. To illustrate the importance of this point, start with an initial level of real aggregate demand that is in equilibrium with aggregate supply. Assume that an expansionary monetary policy increases aggregate demand at the existing level of prices (and nominal wages), so that relative prices (and relative and absolute real wages) are also at their initial levels. The increase in aggregate demand increases the demand for each commodity. Further, assume that the price-setting firm has increasing marginal cost. The response to the increase in aggregate demand would then be as follows.

- If all firms adjusted their prices simultaneously and continuously, all prices would rise at the same time and relative prices would not change, so that individual firms would not have an incentive to increase output. Further, the price level would rise in the same proportion as the initial increase in aggregate demand, so that aggregate demand at the now higher price level would revert to its initial real value and demand pressure on prices would be eliminated. Therefore, monetary policy would be neutral. If the initial increase in aggregate demand were due to an increase in the money supply, the real value of the money supply would revert to its initial level. If the initial increase in aggregate demand were due to a central bank's action to lower the interest rate, the price level would continue to rise until the central bank reverses its action and raises the interest rate (see Wicksell's pure credit economy analysis in Chapter 2), which is needed to return investment and consumption to their initial levels.
- If individual prices are adjusted in a staggered (rather than simultaneous) manner and discretely, rather than continuously, as in the sticky price theory, there would be slow adjustment of individual prices[23] and hence of the price level, so that the real value of aggregate demand would remain higher than in the initial equilibrium and the firms would be producing greater output. Hence, an expansionary monetary policy would have resulted in greater output during the adjustment process and would not be neutral.
- In the long run, once all adjustments are completed, all prices would have adjusted, so that relative prices and output would be as in the initial equilibrium. The price level would have increased in proportion to the increase in aggregate demand and monetary policy would become neutral.
- For a given increase in demand and, therefore, in the marginal revenue of the monopolistically competitive firm, the speed at which the price level adjusts to its long-run value depends on the elasticity of marginal cost. For a given increase in demand, the flatter the firm's marginal cost curve and the smaller the increase in the firm's price, the greater the increase in output[24] and the slower the return to long-run

23 This process has been likened to the movement of a chain gang (whose members are tied together by a chain). Usually, the larger the number of members of the chain gang and/or the shorter the chain, the more slowly would the chain gang tend to move.

24 This argument is sometimes stated as follows. For monopolistic firms, the profit-maximizing intersection of the marginal cost and marginal revenue curves is below the firm's price, so that small increases in demand will be accommodated by an output increase as long as marginal cost remains below this price. However larger demand increases that push marginal cost above the price will cause the firm to raise its price, though this is accompanied by a relatively smaller or zero output increase.

> monetary neutrality. In the context of monopolistic competition (which itself is a "small departure" from the perfectly competitive economy) and interpreting the low elasticity of marginal cost and therefore of the firm's relative price as a "real rigidity," this point is sometimes stated as the proposition: in response to changes in aggregate demand, small real rigidities can generate substantial "nominal rigidity" of the price level, which, in turn, can cause substantial departures from the neutrality of monetary policy (Blanchard, 2000, pp. 1390–91).

Note that the above conclusions pertain to monopolistic competition in commodity markets. The assumption of monopolistic competition seems to be less realistic for labor markets, and may or may not be essential to support the preceding results. However, labor markets are heterogeneous and also have departures from perfect competition and "rigidities." At least some of them are encompassed in Keynes's arguments about labor markets and the more recent implicit contract and efficiency wage theories, as well as in the theory of staggered nominal wage contracts under uncertainty (see Chapter 14 on the expectations-augmented Phillips curve).

15.4 New Keynesian (NK) macroeconomics

The new Keynesian (NK) school in macroeconomics emerged in the 1990s (Clarida, Gali and Gertler (CGG), 1999; Gali, 2002). Its general philosophy is in the Keynesian tradition and its major components are from the neoKeynesian collection of ideas. While the latter was a disparate rather than a tightly specified macroeconomic model, the NK school offers a tightly specified, integrated macroeconomic model that rebottles and re-flavors some of the neoKeynesian ideas, such as that of sticky prices, while abandoning or ignoring others, such as the efficiency wage theory and nominal wage theory. It has also adopted the Taylor rule (see Chapter 13) for the formulation of monetary policy. Its major departure from previous Keynesian economics comes from its methodology. While the methodology of earlier Keynesian thought had been somewhat eclectic and had usually used one-period comparative static analysis, the new Keynesian school adopts for its methodology stochastic, intertemporal optimization and market clearance, which are the hallmark of the modern classical school, especially its real business cycle theory. It also adopts the latter's rational expectations hypothesis (REH). For its distinctive results, the new Keynesian school relies on staggered price adjustments by monopolistically competitive firms and the assumption that the central bank sets the interest rate, not the money supply, and does it through a forward-looking Taylor rule derived from optimization of the objective function of the central bank. The resulting model has become known as the NK model. Its use of the interest rate, with the money supply made endogenous, as the critical determinant of aggregate demand, is in the tradition of Wicksell's model for the pure credit economy (see Chapter 2), so that it has sometimes been called a "neo-Wicksellian" model.

While the standard NK model, as presented below, is based on imperfections in commodity markets, it assumes that the markets for non-monetary financial assets are complete and work perfectly, with perfect substitution among all non-monetary financial assets, so that the distinction between bonds and loans/credit is irrelevant. However, another innovation by Keynesians during the last two decades has emphasized imperfect substitution between these financial assets, leading to the extension of macroeconomic modeling to incorporate credit as a separate asset from money and bonds. The distinction between these two types of assets and its implications for monetary policy are presented in the next chapter.

The following subsections present the main new Keynesian ideas on the commodity and labor markets and on monetary policy. For the formal model, it relies on CGG (Clarida *et al*., 1999, Ch. 5). Walsh (2003) presents an in-depth review of new Keynesian models.

15.4.1 NK commodity market analysis

The NK closed economy (expectational) IS equation for equilibrium in the commodity market is:

$$y_t = c_t + i_t + g_t$$

where c is real consumption, i is real investment and g is real government expenditures on commodities. In the NK model, both households and firms pursue intertemporal optimization and use rational expectations to predict the future values of the relevant variables. Intertemporal optimization by households implies that current consumption will depend on the real interest rate and both the current and future levels of actual output (not the long-run equilibrium level), as in the life-cycle and permanent income hypotheses of consumption. Intertemporal optimization by firms implies that current investment will depend on the real interest rate and the future desired capital stock, which will depend on the future demand for commodities. Consequently, the current demand for commodities will depend on the future demand for commodities by households and firms and the real interest rate. Hence, the general form of the NK IS equation for equilibrium in the commodity market is:

$$y_t = f(y_{t+1}, r_t, g_t)$$

where g is taken to represent all sources of expenditures other than consumption and investment, as well as demand shocks. The NK model of CGG states this closed-economy "expectational IS relationship" for equilibrium in the commodity market as:

$$x_t = f(x_{t+1}, r_t, g_t)$$

where $x_t = y_t - y^{\mathrm{f}}$, so that x represents the output gap. Assuming rational expectations and the Fisher equation, CGG approximate the preceding relationship by the log-linear expression:

$$x_t = E_t(x_{t+1}) - \psi(R_t - E_t\pi_{t+1}) + g_t \tag{13}$$

Further, CGG assume that g follows a first-order autoregressive process, given by:

$$g_t = \alpha g_{t-1} + \mu_t$$

where μ is a random variable with zero mean and constant variance. Note that, since both consumption and investment are (only) forward looking and the rational expectations hypothesis is applied to their forward values, the past levels of expenditures do not affect their current levels.[25] Only the expectations of their future levels do so. Therefore, persistence

25 Habit persistence in consumption is excluded from the model, as is the impact of past incomes through their impact on the inherited capital stock.

(i.e. the impact of the past on the present) had to be introduced into the IS equation through the specification of g.

Note that iterating (13) forward yields the NK IS equation as:

$$x_t = E_t \left[\sum_{j=0}^{\infty} -\Psi(R_{t+j} - \pi_{t+1+j} + g_{t+j} \right] \tag{14}$$

15.4.2 NK price adjustment analysis

For the price adjustment process, the new Keynesian school assumes monopolistic competition in commodity markets and that each firm sets its product price so as to maximize its profits subject to the cost and frequency of expected future price adjustments. The latter implies that the firms set prices as in the menu-cost theory (presented earlier), with staggered price adjustments at different times by different firms in the economy.[26] However, the NK literature specifies the firm's price adjustment using a time-contingent pattern specified by Calvo (1983). Under this pattern, the representative firm decides on the probability θ that it will keep its price fixed this period, so that the probability that it will adjust its price is $(1-\theta)$.[27] θ is assumed to be independent of the time when the firm last adjusted its price, so that the adjustment made this period is independent of the past history of adjustments. Hence, the average time for which the representative firm's price remains fixed is $1/(1-\theta)$.[28] However, θ will depend on the expected future price adjustments, as will the current price adjustment.

The profit-maximizing price adjustment by a price-setting firm will depend on its marginal cost. This marginal cost rises with the level of output and the prices of inputs, which are proxied by the price level. Therefore, in logs, the representative firm's desired price p^*_t for its product can be expressed as:

$$p^*_t = P_t + \alpha x_t \tag{15}$$

where P is the (log of the) price level, taken to proxy input costs, and x, as before, is the (log of the) representative firm's share of the deviation of output from its full-employment level, with α representing the responsiveness of desired prices to the level of activity x. However, because price adjustment is only periodic, the price p_t to which the firm adjusts in the current period t is determined by a weighted average of the current and

26 Eichenbaum and Fisher (2007) estimate, for their revision of the imperfect competition model (to allow variable elasticity of demand and firm-specific capital), that firms re-optimize prices on average every two quarters, rather than every two years, which implies a high degree of price inertia in the version without these modifications. They also find that the Calvo-style models of only periodic price adjustments can account for the behavior of post-war inflation in the USA.

27 Alternatively, the proportion $(1-\theta)$ of firms can be assumed to change their prices in any given period, while the remainder do not do so. In this case, each firm has the same probability, $(1-\theta)$, of being a price-adjusting firm, irrespective of when it last adjusted its price. Some economists have suggested that θ should be determined endogenously and allowed to vary.

28 In a quarterly model, $\theta = 0.75$ implies that the price is adjusted on average once a year.

future expected price adjustments, so that:

$$p_t = \lambda \sum_{j=0}^{\infty} (1-\lambda)^j E_t p^*_{t+j} \qquad 0 < \lambda \leq 1 \tag{16}$$[29]

where λ is the rate at which price adjustments will be made.

The price level is the average of all prices in the economy, so that it is a weighted average of all the prices adjusted in the past. Hence:

$$P_t = \lambda \sum_{j=0}^{\infty} (1-\lambda)^j E_t p_{t-j} \tag{17}$$

Rewrite (16) and (17) as:

$$p_t = \lambda p^*{}_t + (1-\lambda) E_t p_{t+1} \tag{18}$$

$$P_t = \lambda p_t + (1-\lambda) P_{t-1} \tag{19}$$

From (15), (18) and (19), and using $\pi_t = P_t - P_{t-1}$, we get:

$$\pi_t = E_t \pi_{t+1} + [\alpha \lambda^2/(1-\lambda)] x_t \tag{20}$$[30]

Hence, current inflation is a function of the current output gap and next period's expected inflation rate. Note that λ is the frequency of price adjustments and α is the responsiveness of the firm's desired price to the level of output relative to its full-employment level. Lower values of λ and α mean less responsiveness of inflation to current activity levels. (20) implies that higher future levels of inflation will raise the current inflation rate. Further, the lower the production relative to its full employment level, the lower will be the current inflation.

CGG state the economy's short-run supply or price/inflation adjustment equation as:

$$\pi_t = \gamma x_t + \beta E_t \pi_{t+1} + z_t \tag{21}$$

where x is now the deviation of the economy's output from the full-employment level and z_t represents shocks, such as to the monopoly markup, that affect marginal cost; z_t is sometimes referred to as the "cost-push" element of inflation. γ is a decreasing function of θ, so that the longer the price remains sticky or unchanged, the less the elasticity of inflation to the output gap. CGG specify z_t by:

$$z_t = \rho z_{t-1} + v_t$$

29 Desired prices further in the future have a lower weight because the possibility of price adjustments before that date is greater.

30 The price adjustment process can also be derived from the following intuitive argument. π_t is a function of $\pi^e{}_{t+1}$ and mc_t, where mc is marginal cost, which affects price-adjustments since profit-maximizing firms want to equate marginal revenue to marginal cost. Future inflation affects current inflation because firms smoothen price changes to reduce the costs of changing prices. Since marginal cost mc rises with an increase in output, it depends on the output gap x. Replacing mc by x, we get $\pi_t = f(\pi^e{}_{t+1}, x_t)$.

where v is a random variable with zero mean and constant variance,[31] so that z_t follows a stochastic first-order regressive process.

In terms of the impact of monetary policy on inflation, prices adjust gradually to the price level that would have occurred if the economy had remained at full employment. If firms expect a lower money supply in the future, they will also expect lower prices in the future. Their optimal response is to start by lowering current prices, which increases current real-money supply and output. Hence, the forward-looking response to future money supply shocks is to smoothen inflation over time but to make current output a negative function of future money supply changes.

New Keynesian economics considers the price adjustment equation (21) as its version of the Phillips curve, since it relates the inflation rate to the output gap, so that it is often referred to as the new Keynesian Phillips curve (NKPC). However, the original Phillips curve reflected labor market behavior, whereas the usual NK price adjustment analysis does not incorporate an explicit analysis of the labor market, though one can be appended to the NK model. This represents a glaring omission of the NK model, which is especially striking given the role assigned to labor markets and their imperfect functioning in Keynes's own analysis and those of the Keynesian models preceding the NK model. Since the analytical basis and form of (20) are derived from the optimal response to changes in demand along the marginal cost curve of the firm, rather than from the behavior of labor markets, calling it a Phillips curve is a misnomer.[32] The CGG version of the NK model does not explicitly specify labor demand and supply, and does not lay out the process for the determination of or changes in nominal or real wages. Given this, the only explicit basis for price adjustment and inflation comes from the positive slope of the marginal cost curve and the gradual adjustment of prices. Variations in work effort, as in the implicit contract and labor hoarding theory, which would produce shifts of the marginal cost curve over the business cycle, are also neglected.[33]

Further, note that given the usual upward slope of the marginal cost curve, output changes cannot occur unless they are accompanied by price changes. Therefore, (20) cannot explain how monetary policy can, at least sometimes, change output without prior or accompanying changes in prices/inflation.

The price/inflation adjustment process (20) replaces the short-run aggregate supply function of the AD–AS models, based on the labor market and production analysis in classical models and the traditional Phillips curve in Keynesian models. Iterating (21) forward yields:

$$\pi_t = E_t \left[\sum_{j=0}^{\infty} \beta^j (\lambda x_{t+j} + z_{t+j}) \right] \tag{22}$$

31 z can be interpreted as representing deviations from a linear impact of the output gap on marginal cost, but it can also encompass other sources of deviations. See Gali and Gertler (1999), Clarida *et al*. (CGG) (1999, page 1667, footnote 15).

32 Equation (20) is also different from the expectations-augmented Phillips curve (EAPC) since its right-hand side has the inflation rate expected for future periods, whereas the EAPC referred to the deviation of the current inflation rate from its own expected level.

33 Further, since the basis for inflation lies in the positive slope of the marginal cost curve, inflation could not be explained by the output gap if marginal cost was constant with respect to the output gap. This is especially likely to be so given variations in labor effort and labor hoarding.

so that the current inflation rate depends on the current and future output gaps: firms set current prices on the basis of the expectation of future marginal costs, which vary with future demand relative to the full-employment output.

Long-run supply function in the NK model

The long-run commodity supply function for the new Keynesian model, as for the modern classical model, is based on the assumptions of perfect price and wage flexibility, absence of adjustment costs, and is derived from intertemporal optimization by firms and consumers/workers. This long-run commodity supply is at the full-employment level and is independent of aggregate demand and its determinants, so that the long-run value of the output gap, x^{LR}, is zero. In addition, in (21), with stochastic errors set at zero for the long-run analysis, $z_t = z_{t-1}$. Hence, for the long run, the price adjustment equation becomes:

$$\pi^{\mathrm{LR}}{}_t = \beta E_t \pi^{\mathrm{LR}}{}_{t+1} + v_t \tag{23}$$

which makes current inflation a function only of future expected inflation, rather than of current aggregate demand and supply.

15.4.3 Other reasons for sticky prices, output and employment

The NK sticky information hypothesis

To get around the invalidity of the forward-looking NKPC based on sticky prices and to introduce inflation inertia, Mankiw and Reis (2002, 2006a,b) propose a staggered "sticky information" hypothesis as a basis for a price adjustment equation. Under this hypothesis, information is costly to acquire and process, so that it is updated only periodically. Adopting the Calvo process to staggered information, a fraction λ of the firms update their information each period and adjust their price, while $(1 - \lambda)$ are "inattentive." The adjusting firms are again drawn randomly from all firms, so that the current price level is a weighted average of past prices, rather than of the future expected price.

$$P_t = \lambda \sum_{j=0}^{\infty} (1-\lambda)^j P_{t-j} \qquad 0 < \lambda \leq 1 \tag{24}$$[34]

We do not specify the Mankiw and Reis model further and leave it to the reader. Suffice it to say that the price adjustment process in (24) is backward looking and generates persistence in inflation, unlike the sticky prices process. Similar sticky information processes can also be attributed to consumers and workers. While this adjustment process is closer to the traditional Phillips curve or one with static or adaptive expectations, Mankiw and Reis argue that the sticky information hypothesis provides a preferable foundation for inflation inertia in terms of microeconomic optimization.

34 In the determination of the current price level, price levels further in the past would have relatively lower weights since they would also be incorporated in more recent price levels.

Other reasons for staggered and hesitant price and production strategies

One source of staggered price and output adjustments in response to a shift in demand or supply functions is the hypothesis that the monopolistically competitive firm faces adjustment costs of changing prices, output and employment. For any one of these three variables, the simplest such hypothesis posits a quadratic cost function for the adjustment from last period's level of the variable, whose minimization implies that it will adjust partially each period[35] (see Chapter 8 for an analysis based on adjustment costs). Aggregating the adjustment cost of all three variables, cost minimization will imply the partial adjustment each period of the firm's product price, employment and output, with the changes made becoming a function of last period's price and output, so that there will occur short-run persistence, with decay, in output, employment and price variations over time.

Another source of staggered adjustments by the firm arises from the nature of the flow of information and risk aversion by firms. Leaving aside the possibility, considered above, of information being sticky, information is not only costly to process, it arrives in a discrete, staggered manner and is usually inadequate and ambiguous and often contains contradictory signals. In terms of its signal, such information is not only incomplete and fuzzy (meaning vague) but also "dirty," by which we mean that different parts of it provide signals contrary to those coming from other parts, so that the overall message is not transparent. To illustrate, suppose that the firm relies on a number of leading indicators of the economy's aggregate demand and output levels in predicting the demand for its product, these being of the type commonly used by the central bank and economic analysts. Usually, at a given time, while incoming data on some leading indicators of real output might signal a future decline in aggregate demand and output, the data on some others would point to an increase, while the relevant data on many other indicators would still not be available. Hence, altogether, the information is not only incomplete but its overall signal is fuzzy and dirty, so that there is "fundamental uncertainty" rather than merely knowable, risk-based information.[36] In such a context, suppose that the firm forms a subjective probability distribution on the change in economic activity in the quarter ahead, with the expected mean being a negative one. Being risk averse, it reduces its own output and price not by the amount that it would do if the standard deviation of the distribution was zero – that is, it was sure of a decline and of its amount – but by a lesser amount. Further, given the fuzziness and dirtiness of the signal,[37] the firm would be averse to adopting measures that would be more costly to reverse. The decline in the risk-averse firm's production may then be achieved not by laying off a sufficient number of workers, but by reducing their effort, as explained by the implicit contract theory, and any cut in its product price may be implemented by introducing discounts and special offers, rather than by a published cut in its product price. If the information over the quarter does not change, and the expectation of a decline becomes firmer, it could follow through in the following quarter with another reduction. But if the new information indicates a movement

35 Among other studies with such an assumption, Ireland (2001) reports support for the assumption that firms face a quadratic cost of price adjustment rather than inflation adjustment.

36 Post-Keynesians emphasized the fundamental nature of uncertainty as arising from inadequate, internally conflicting and staggered arrival of information, whose overall assessment could shift with each new bit of information. By comparison, the modern classical, and even the new Keynesian, models assume knowable, internally consistent and adequate information.

37 Under this nature of the available information, news and rumors can play a significant part in revisions of expectations and decisions.

of the distribution to a smaller decline in economic activity than had seemed earlier, it may decide not to undertake a further reduction. However, if the revision in information indicates a pick-up in economic activity, so that the earlier assessment is now itself assessed to be incorrect, the firm can easily reverse the reduction in its own price and production. Hence, the firm's usual risk-averse strategy amounts to its pursuing hesitant, staggered price and production strategies in small steps over time. Given the collection of firms, with different sources of data (e.g. because they are in different industries) and different assessments of the subjective probability distributions, the economy as a whole would have undergone staggered and gradual price and production adjustments.

Note that this mode of information and action is no different in nature from that of the central bank, which, even with all the data and information at its disposal in modern economies, still finds the incoming information at any given point in time to be incomplete, fuzzy and dirty. Its use of its policy instrument, whether the money supply or the interest rate, also tends to be hesitant, in small steps, and often leaves the door open for no further change, another change in the same direction or a reversal of the previous change.

To conclude, there can be many sources of stickiness of prices, output and employment. Collectively, they provide a very robust basis for a gradual adjustment pattern in these variables, as opposed to complete and instantaneous adjustments, in response to demand and supply shifts.

15.4.4 Interest rate determination

New Keynesian economics has adopted the currently popular assumption that the central bank uses the interest rate, rather than money supply targeting, as its main instrument of monetary policy and acts as if it decides on the interest rate through a Taylor rule.[38] With the interest rate as the primary monetary policy instrument, the central bank adjusts the monetary base to ensure equilibrium in the money and other financial markets, so that the money supply becomes an endogenous variable and is no longer relevant to the determination of aggregate demand, output, employment, the price level or inflation. It can thus be removed from the macroeconomic analysis for the determination of these variables. As argued in Chapter 13, the endogenous determination of the money supply by money demand makes the LM equation and LM curve irrelevant to macroeconomic modeling.

Different forms of the Taylor rule

As discussed in Chapter 13, the Taylor rule is a feedback rule that makes the economy's interest rate a function of the output gap and the deviation of the inflation rate from the target rate, so that the interest rate responds to deviations of output and inflation from their long-run values. There are three basic forms of this rule, as follows.

Contemporaneous Taylor rule (Taylor, 1993):

$$r^{\mathrm{T}}{}_t = r_0 + \alpha x_t + \beta(\pi_t - \pi^{\mathrm{T}}) \qquad \alpha, \beta > 0 \tag{25}$$

38 This is supported, among others, by Taylor (1993), Rudebusch (1995) and Clarida *et al*. (1999, 2000). Levin *et al*. (1999, 2001) report for US data that a simple version of the inflation and output-targeting rule for the US economy does quite well and that responding to an inflation forecast, not longer than a year, performs better than forecasts of inflation farther into the future.

Backward-looking Taylor rule:

$$r^{\mathrm{T}}{}_t = r_0 + \alpha x_t + \beta(\pi_{t-1} - \pi^{\mathrm{T}}) \qquad \alpha, \beta > 0 \tag{26}$$

Forward-looking Taylor rule:

$$r^{\mathrm{T}}{}_t = r_0 + \alpha E_t x_{t+1} + \beta(E_t \pi_{t+1} - \pi^{\mathrm{T}}) \qquad \alpha, \beta > 0 \tag{27}$$

In these rules, r_0 is often taken to represent the long-run interest rate. However, since the central bank's current decision is usually relative to the previous period's interest rate, some studies modify the Taylor rules to incorporate the last period's interest rate and, in order to introduce interest rate smoothing,[39] sometimes also the deviation Dr_{t-1} of the last period's interest rate from its long-run equilibrium level. Doing both of these modifies the above rules to the following ones.

Contemporaneous Taylor rule with persistence/smoothing:

$$r^{\mathrm{T}}{}_t = r_{t-1} + \lambda Dr_{t-1} + (1-\lambda)[\alpha x_t + \beta(\pi_t - \pi^{\mathrm{T}})] \qquad \alpha, \beta > 0, 0 \leq \lambda \leq 1 \tag{28}$$

Backward-looking Taylor rule with persistence/smoothing:

$$r^{\mathrm{T}}{}_t = r_{t-1} + \lambda Dr_{t-1} + (1-\lambda)[\alpha x_t + \beta(\pi_{t-1} - \pi^{\mathrm{T}})] \qquad \alpha, \beta > 0, 0 \leq \lambda \leq 1 \tag{29}$$

Forward-looking Taylor rule with persistence/smoothing:

$$r^{\mathrm{T}}{}_t = r_{t-1} + \lambda Dr_{t-1} + (1-\lambda)[\alpha E_t x_{t+1} + \beta(E_t \pi_{t+1} - \pi^{\mathrm{T}})] \qquad \alpha, \beta > 0, 0 \leq \lambda \leq 1 \tag{30}$$

In these rules, $\mathrm{D}r_{t-1}$ is the deviation of the last period's interest rate from its long-run equilibrium level. The backward-looking (forward-looking) rule can be augmented to include additional backward (forward) output and inflation gaps.

Because of lags in the impact of monetary policy, a forward-looking Taylor rule is preferable to the contemporaneous and backward-looking ones. There is almost always a time lag, estimated at about six quarters or so for many Western economies, between the change in the interest rate and its impact on inflation and the output gap, so that current monetary policy should be formulated to address future inflationary pressures. If interest rates were increased in response to current inflation, their impact would occur at a future date when inflation is likely to have become different from the current rate, so that the policy could have an undesired impact. Although it is difficult to accurately predict future inflationary pressures, it may still be preferable to use a forward-looking Taylor rule.

Note that the convexity of the Phillips curve implies that inflation will react more strongly, and output less strongly, to a positive increase in aggregate demand than they would do to a corresponding decrease in aggregate demand: in absolute terms, a given increase in demand will increase inflation more than the decrease in inflation caused by a corresponding decrease in demand. Therefore, in order for the Taylor rule to reflect the asymmetric effects of aggregate demand changes on inflation and output, it too would have to be asymmetric.

39 A simple modification of the Taylor rule to allow interest rate smoothing does so by introducing the adjustment pattern: $r^{\mathrm{T}}{}_t = \rho r_{t-1} + (1-\rho) r^{\mathrm{T}}$, $0 \leq \rho \leq 1$.

The form of the Taylor rule can be specified a priori or on an empirical basis, or through optimization of the central bank's objective function. The empirical basis relies on estimations, which usually favor a backward-looking rule. However, there are also two other ways of deriving the appropriate form of the Taylor rule. One of these comes from the argument that interest rate changes usually start to impact output and inflation only after several quarters and then continue to do so for several more quarters. If the central bank wants to set the current interest rate to address future output and inflation gaps, the appropriate form of the Taylor rule would have to be forward looking but with a distributed lag pattern over several quarters, quite possibly more than eight. The second form of the appropriate version of the Taylor rule is derived by optimization of a welfare loss function. The NK model takes this route. This is presented in the next section.

New Keynesian derivation of the forward-looking Taylor rule

While the Taylor rule can be stated as an institutional datum, some new Keynesians (CGG, 1999) prefer to derive it from the central bank's objective function. For this, the objective function is taken to be an intertemporal one over current and future output gaps and inflation, and is usually specified as the negative of a quadratic "loss function," so that it has the form:

$$-\frac{1}{2}E_t\left\{\sum_{j=0}^{\infty}\beta^j\left(\gamma x_{t+j}^2+\left(\pi_{t+j}-\pi^{\mathrm{T}}\right)^2\right)\right\} \qquad (31)^{40}$$

where x is the output gap, so that the target output level is assumed to be the full-employment level. π^{T} is the target inflation rate, β is the central bank's time discount factor[41] and γ is the weight on the output gap relative to that on inflation (or the "inflation gap," defined as the difference between the actual inflation rate and the target one). γ is a function of the preference and technology parameters. If the target inflation rate generates the actual trend, the inflation gap will also represent the deviation of current inflation from the trend.

The central bank maximizes (with respect to x and π) the above objective function (i.e. minimizes the quadratic loss function) subject to the linear constraints specified by the IS equation and the price adjustment equation imposed by the economy (CGG, 1999). Doing so for the optimal or target real interest rate, and using $E_t\pi_{t+1}$ to represent the rationally expected future levels of inflation, yields the target interest rate as:

$$r^{\mathrm{T}}{}_t=\alpha+\lambda_x x_t+\lambda_\pi(E_t\pi_{t+1}-\pi^{\mathrm{T}}) \qquad \lambda_x,\lambda_y>0 \qquad (32)$$

In this derivation of the central bank's policy rule, the policy responds to the expected future inflation rate rather than to the current one, unlike that in the standard Taylor rule. It is, therefore, a "forward looking" version of the Taylor rule and can be labeled the NK Taylor rule. In the long run, since the economy functions at full employment, $x^{\mathrm{LR}}=0$.

40 Note, for comparison, that Rotemberg and Woodford (1999) base the central bank's objective function on the representative individual's welfare function. In this function, the deviation of inflation from its trend (rather than the inflation rate itself) has negative utility because this deviation makes it more difficult for economic agents to plan for consumption, investment and portfolio allocations. The individual's welfare also depends on the output gap since this gap is associated with fluctuations in employment and income.

41 Some studies identify this discount factor as the representative household's one.

Further, in the long run, since the central bank is taken to be able to achieve its target inflation rate, $(E_t\pi_{t+1} - \pi^{\mathrm{T}})^{\mathrm{LR}} = 0$. Therefore, $\alpha = r^{\mathrm{LR}}$. Hence, the CGG "forward-looking Taylor rule" becomes:

$$r^{\mathrm{T}}{}_t = r^{\mathrm{LR}} + \lambda_x X_t + \lambda_\pi (E_t\pi_{t+1} - \pi^{\mathrm{T}}) \qquad \lambda_x, \lambda_\pi > 0 \tag{33}$$

where r^{LR} is determined by the long-run supply function for commodities and the IS equation (13). Since perfect capital markets are being assumed, the long-run nominal interest rate R^{LR}, with $\pi^{\mathrm{e}} = \pi^{\mathrm{LR}} = \pi^{\mathrm{T}}$, is given by the Fisher equation as:

$$R^{\mathrm{LR}} = r^{\mathrm{LR}} + \pi^{\mathrm{T}}$$

Further, CGG (1999) allow for interest rate smoothing by letting the actual real interest rate be set by the central bank as:

$$r_t = \rho r_{t-1} + (1 - \rho) r^{\mathrm{T}}{}_{t-1} \tag{34}$$

where $r^{\mathrm{T}}{}_{t-1}$ is specified by lagging (33).

Monetary policy in the NK model

The CGG version of the Taylor rule implies that the central bank should raise the real interest rate if a positive shock to aggregate demand increases the inflation rate above its target level and/or output above its full-employment level. Hence, since the central bank usually manipulates the nominal rate in the financial markets, it should raise the nominal rate more than the inflation rate.

For positive permanent increases in the aggregate supply of commodities, CGG assume that such a policy would raise permanent income and increase aggregate demand to the same extent, so that the output gap will not change. Nor will there be any pressure on inflation. Therefore, monetary policy would not have to respond to permanent supply shocks. However, the required assumption for this result is that the time path of the increase in aggregate demand will match that in aggregate supply.[42] While consumption does respond to permanent income, the short-run marginal propensity to consume out of permanent income is less than unity, so that consumption will rise by less than aggregate supply. If the other components of aggregate demand do not rise by exactly enough to fill the gap between increasing aggregate supply and increasing consumption, aggregate demand will increase less than aggregate supply. In this eventuality, output demand and short-run output will fall below the full-employment levels, creating a negative output gap, and will also lower inflation below its target level. Consequently, the Taylor rule will imply that the real interest rate will have to be reduced as a short-run measure. As against this tendency, the increase in the productivity of capital accompanying the increase in full-employment output may induce firms to increase investment, which will increase aggregate demand. Further, positive shocks to output are

42 This evokes memories of Say's law, which was the assertion that increases in full-employment output will automatically increase demand to the same extent (see Chapter 18). Keynes made the objection to such an automatic response the cornerstone of his economics, and Keynesians usually adhere to it. It should not be made, explicitly or implicitly, a part of new Keynesian economics.

sometimes accompanied by increases in consumer and business confidence, with euphoria and exuberance, which can raise stock prices, investment and consumption, thereby increasing aggregate demand. Hence, there is no guarantee that aggregate supply shifts will, or usually do, increase aggregate demand to the same extent. Therefore, the short-run policy response under the Taylor rule cannot be predicted a priori but will depend on the actual increase in aggregate demand relative to the increase in aggregate supply. Hence, monetary policy will have to respond in an appropriate manner to shifts in supply.

Although NK models rely on sticky prices, they, just like the Friedman–Lucas supply rule in the absence of sticky prices, imply that the central bank can reduce inflation without any cost in terms of output as long as expectations of inflation change at the same rate as inflation itself. The latter would require the pursuit of a credible policy, announced sufficiently in advance to allow firms to change their price adjustments. If the revision in policies from past patterns of higher inflation is not sufficiently credible or is not announced in sufficient time, disinflation will impose reductions in output.

15.4.5 Variations of the overall NK model

The NK model has several versions, with their common feature being that there are three equations: an IS-type equation for commodity market equilibrium, a Taylor rule for the interest rate and an aggregate supply or price-adjustment (the "NK Phillips" curve) equation. To illustrate this diversity of versions, we specify below Woodford's (2007) version of the NK model.

Woodford's specification of what he calls an "intertemporal IS relation," derived from an Euler equation of the timing of aggregate expenditures, is:

$$\ln(y_t/y_t^{\mathrm{f}}) = E_t[\ln(y_{t+1}/y_{t+1}^{\mathrm{f}}) - \sigma[R_t - E_t\pi_{t+1} - r_t^{\mathrm{n}}] \tag{35}$$

where y is (real) output of commodities, y^{f} is its long-run equilibrium (full-employment) level;[43] R is the one-period nominal interest rate on riskless assets, r is the real interest rate and r^{n} is its long-run equilibrium (full-employment) value; and π is the inflation rate. Future values of the variables are rationally expected ones. There is a strong intertemporal element embodied in this equation. Compared with the usual IS equation that has y on the left-hand side, the preceding equation is specified in the form of the output gap $y_{t+1}/y_{t+1}^{\mathrm{f}}$ to facilitate the solution of the model.

The aggregate supply or price/inflation adjustment equation is:

$$\pi_t - \pi^*{}_t = \alpha \ln(y_t/y_t^{\mathrm{f}}) + \beta E_t(\pi_{t+1} - \pi^*{}_{t+1}) + \mu_t \qquad \alpha > 0, 0 < \beta < 1 \tag{36}$$

where $\pi^*{}_{t+1}$ is the perceived or expected rate of *trend* inflation at time t.[44] The disturbance term μ incorporates exogenous cost-push elements. This equation is interpreted as a log-linear approximation of the staggered price dynamics in Calvo (1983). In this equation, if firms do not reoptimize their prices in a period, they automatically increase them at the trend inflation rate π^*, so that a change in their relative price only comes about through reoptimization.

43 The long-run equilibrium values are determined by exogenous real factors such as production technology, population and household preferences.

44 If economic agents expect that the central bank will achieve its target for the trend inflation rate, $\pi^*{}_{t+1}$ will equal this target trend rate.

The third equation is the monetary policy one, specified by:

$$R_t = r^*_t + \pi^*_t + \lambda_\pi(\pi_t - \pi^*_t) + \lambda_y \ln(y_t/y_t^{\mathrm{f}}) \tag{37}$$

where π^*_t is the central bank's inflation target for period t and r^*_t is the bank's perceived value of the economy's equilibrium real interest rate in period t. Both these variables are taken to be exogenous, with shifts in them reflecting shifts in attitudes within the central bank. It is further assumed that π^*_t follows a random walk with mean zero, so that:

$$\pi^*_t = \pi^*_{t-1} + v_t^{\pi} \tag{38}$$

The complete model and its derivation and implications can be found in Woodford (2007, pp. 3–8).

15.4.6 Money supply in the NK model

The NK model does not explicitly have money supply in any of its (three) core equations, even though they determine the price level and the inflation rate. It would therefore seem that its determination of these variables is independent of the money supply and its growth rate, so that the central bank could ignore monetary aggregates and the money demand function altogether. It would also seem from its price adjustment equation, derived from firm's profit maximization, that the price level and the inflation rate are determined by the actions of firms in setting relative prices. These conclusions would be erroneous for the NK model.

Monetary policy in the NK model is determined by the central bank through its pursuit of a Taylor-type interest rate rule. Chapter 13 had argued that, if the central bank wants to achieve the interest rate that it sets (under this rule or otherwise), it must ensure the appropriate money supply in the economy. It had also argued that this money supply would be the one that ensures equilibrium in the money market, so that it must equal the money demanded at the set interest rate and the economy's values of the other determinants of money demand. To illustrate, assume that the money demand function has the linear form:

$$m^{\mathrm{d}} = m^{\mathrm{d}}(y^{\mathrm{d}}, R, FW_0) = m_y y^{\mathrm{d}} + (FW_0 - m_R R) + \eta_t \tag{39}$$

where $0 < m_y \leq 1$, $0 < m_R < \infty$ and η_t is a disturbance term. The definitions of the symbols in this and the following equations are as given in Chapter 13. For perfect capital markets, the Fisher equation on interest rates is:

$$R = r + \pi^{\mathrm{e}} + v_t \tag{40}$$

where v_t is a disturbance term. Under a real interest rate rule, r is set at r^{T} by the central bank and y^{d} is determined above by the AD equation. Using, for illustration, the linear commodity sector model specified in Chapter 13 for the open economy, the AD equation is:

$$y^{\mathrm{d}} = \alpha[\{c_0 - c_y t_0 + i_0 - i_r r^{\mathrm{T}} + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} + (1/\rho^{\mathrm{r}})\cdot\{-z_{c0} + z_{cy}t_0 - z_{c\rho}\rho^{\mathrm{r}}\}] + \mu_t \tag{41}$$

where μ_t is a disturbance term. Substituting these values of r^{T} and y^{d} in the money demand equation, we get:

$$\begin{aligned} m^{\mathrm{d}} = m_y\alpha[\{c_0 - c_y t_0 + i_0 - i_r r^{\mathrm{T}} + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} \\ + (1/\rho^{\mathrm{r}})\cdot\{-z_{c0} + z_{cy}t_0 - z_{c\rho}\rho^{\mathrm{r}}\}] - m_R r^{\mathrm{T}} - m_R\pi^{\mathrm{e}} + FW_0 \\ + m_y\mu_t + \eta_t - m_R v_t \end{aligned} \tag{42}$$

If we assume the NK Taylor monetary policy function, r^{T} would be replaced by this function. Since the following results do not depend on the form of this policy function, we avoid writing the more complicated equation that would incorporate this function and, instead, proceed with the preceding equation.

The money market equilibrium condition for the central bank to ensure that the financial markets establish an interest rate equal to its desired target rate, is:

$$\begin{aligned} M^{\mathrm{s}}/P = m_y\alpha[\{c_0 - c_y t_0 + i_0 - i_r r_0^{\mathrm{T}} + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} \\ + (1/\rho^{\mathrm{r}})\cdot\{-z_{c0} + z_{cy}t_0 - z_{c\rho}\rho^{\mathrm{r}}\}] - m_R r^{\mathrm{T}} - m_R\pi^{\mathrm{e}} + FW_0 \\ + m_y\mu_t + \eta_t - m_R v_t \end{aligned} \tag{43}$$

For any given values of P and π^{e}, this equation determines the real money supply M/P required for money market equilibrium. Note that the money market on its own cannot change P, whose movement depends upon the aggregate demand for commodities relative to their supply. Nor can the money market change the money supply M^{s}, which depends upon the monetary base M0 controlled by the central bank, or the monetary base multiplier ($\partial M/\partial$M0), which depends on the payments system and public behavior. Therefore, there is no equilibrating mechanism in the money market that will adjust M/P to real money demand, so that unless the central bank ensures that the economy has the nominal money supply required for money market equilibrium, there is a strong potential for disequilibrium in this market. Such a disequilibrium would mean that the central bank no longer has sufficient control over the market interest rate, on the basis of which the economy determines its aggregate demand, the price level and the inflation rate, so that the central bank would also lose control over these variables.

Therefore, the equilibrium condition for the money market and the ability of the central bank to ensure the required money supply are essential adjuncts of the NK model. Further, the derivations of the values of the market interest rate, the price level, the inflation rate, etc., from the NK model are conditional on the central bank's ability to manage the money supply to achieve equilibrium in the money market.[45] This ability is itself conditional on the economy's money demand function and the stochastic terms.

45 The European Central Bank (ECB) in recent years has espoused what it calls the two pillars of monetary policy. Of these, one is "economic analysis," which assesses the short-to-medium-term determinants of price developments arising from the interplay of the supply and demand for commodities and factors, while the second one is "monetary analysis," which assesses the medium-to-long-term outlook for inflation from the long-run link between money and prices (Woodward, 2007). However, several other banks do not seem to make explicit use of the money supply data. As a result, there is considerable debate in the literature about whether monetary data can contribute to the appropriate formulation of monetary policy and whether it is superfluous or erroneous for the ECB to rely on monetary aggregates. Our analysis indicates a dire need for the use of data on monetary aggregates even in the context of an economy which functions along the lines of the NK model, since the central

Taylor-type money supply rules and the relevance of money supply

If the money demand function were stable and all disturbance terms in (43) were equal to zero, substitution of the Taylor rule for the interest rate in the money market equilibrium condition would yield a money supply function that is itself a type of Taylor rule. For instance, substituting a Taylor interest rate rule of the form:

$$r^{\mathrm{T}}{}_t = r_0 + \alpha(y_t - y^{\mathrm{f}}) + \beta(\pi_t - \pi^{\mathrm{T}}) \quad \alpha, \beta > 0 \tag{44}$$

for the interest rate in the preceding linear money demand function, yields:

$$m^{\mathrm{d}} = m_y y^{\mathrm{d}} + \{FW_0 - m_R \pi^{\mathrm{e}} - m_R(r_0 + \alpha(y_t - y^{\mathrm{f}}) + \beta(\pi_t - \pi^{\mathrm{T}}))\} \tag{45}$$

so that, for money market equilibrium, the money supply rule would be:

$$M^{\mathrm{s}} = P[m_y y^{\mathrm{d}} + \{FW_0 - m_R \pi^{\mathrm{e}} - m_R(r_0 + \alpha(y_t - y^{\mathrm{f}}) + \beta(\pi_t - \pi^{\mathrm{T}}))\}] \tag{46}$$

This equation provides the *money supply rule* that would deliver identical values of the endogenous variables, including the interest rate, under the posited equations for the economy. It is a Taylor-type rule since it makes the money supply a function of the output and inflation "gaps."

However, if the money demand function is not stable and predictable, the preceding money supply rule will have unstable and unknown values of the parameters m_y and m_r, so that the central bank will not know the precise amount of money to supply to the economy. Since this has been the case in recent decades in economies undergoing considerable financial innovation, following a money supply rule has proved to be inferior to following an interest rate rule for controlling aggregate demand.

However, this conclusion need not make the money supply redundant for monetary policy if some of the impact of the money supply on output and inflation were independent of interest rates. For the USA, while Rudebusch and Svensson (2002) find support for the redundancy proposition, Nelson (2002) and Hafer *et al*. (2007) reject such redundancy; money has an impact on output even after controlling for the impact of interest rates on output. Among the reasons offered in Chapter 16 for such a finding is that changes in the money supply could affect the cost and amount of credit, which could have an effect on output other than through interest rates.

Hence, for the NK model, we conclude that, although monetary aggregates do not appear in the NK model's equations, the conduct of monetary policy for this model is not merely the selection and pursuit of a Taylor rule for the target interest rate, it also includes the pursuit of an appropriate policy on the relevant monetary aggregates.[46] This remains so whether the money demand function is stable and known, or not stable and known. The former case

bank needs to ensure the appropriate money supply or face a disconnect between its set interest rate and the market rate.

46 Therefore, it is erroneous to conclude that "money plays no role in the current consensus macroeconomic model and in the conduct of monetary policy." Its erroneous nature can be easily seen if the central bank were to run an experiment that lowers the interest rate but leaves the money supply unchanged or decreases it. In such an experiment, a disconnect would appear between the central bank's target rate and the market rate – and the economy's aggregate demand would evolve on the basis of the market rate.

yields a Taylor-type money supply rule, while the latter does not yield a known target for the money supply or its growth rate, nor a known Taylor-type rule for its determination.

We also note, in passing, that whether the central bank controls the money supply or the interest rate as its policy instrument, it remains pre-eminently the policy authority that can affect the inflation rate and, therefore, is the one to be held accountable for controlling inflation. Correspondingly, its credible commitment to a low inflation target is equally relevant to the achievement of low inflation under both of its operating policy targets (Woodford, 2007).

How does instability of the money demand function affect the provision of the money supply?

The preceding conclusion that the central bank must ensure, through the appropriate money supply, that the economy's interest rate is equal to its desired value holds whether the money demand function is stable or unstable. This task is easier – but still not easy because of the disturbance terms in the money demand function (43) – if the money demand function is stable and known. It is more difficult if the money demand function is unstable and unpredictable, as has been the case in recent decades in many Western economies. Further, if the monetary base multiplier (between the monetary base and the money supply) is also unstable, the achievement of the appropriate money supply through the central bank's control over the monetary base becomes even more problematical.

The central bank can attempt to provide the appropriate money supply through open market operations, reserve requirements, etc. While direct actions of this kind will have to be the main mode of meeting the money supply requirement, Chapter 13 argued that uncertainty over the money demand function and the monetary base multiplier implies that the central bank would have to give the private sector some leeway to bring about changes in the money supply, as, for example, through commercial bank borrowing from the central bank when the market rates rise above the central bank's discount rate.

Monetary aggregates, the quantity equation and inflation

The price level in the NK model is determined by aggregate demand relative to supply, and inflation in this model results from continuous demand versus supply pressures in the commodity market, so that it appears that there is no relation between monetary aggregates and the price level or inflation. Further, monetary aggregates are not in the NK model's three equations. Therefore, it seems as if the links emphasized by the quantity theory and by monetarists between money and prices, or between the money growth rate and inflation, dissolve in the context of an economy operating along the NK lines. Such conclusions are incorrect.

The quantity equation (see Chapter 2) is an identity that must hold in the NK model, as in any other theory. This identity asserts that:

$$P'' \equiv M'' - (y'' - V'') \tag{47}$$

where $''$ stands for the growth rate and V is the velocity of circulation of money. Hence, for given values of y'' and V'', there must be a close relation in a monetary economy between inflation and money growth. The NK model does not – and cannot – repeal the quantity equation. What it can and does do, is to make the money stock endogenous to the interest

rate target and the price level, with the latter determined by the aggregate demand and supply of commodities.[47] Therefore, the more illuminating form of the quantity equation for the NK model is:

$$M'' \equiv P'' + y'' - V'' \tag{48}$$

To conclude, the distinctiveness of the NK model does not lie in that it severs the link between the monetary aggregates and the price level or inflation. Its distinctiveness on this link arises merely from its assumption that the central bank sets an interest rate target, which offers a different route for the determination of aggregate demand than a money supply target policy would. The related aspect of its distinctiveness is that the pursuit by the central bank of the interest rate as its monetary policy instrument makes the money supply endogenous to this interest rate, output and the money demand function. However, the central bank cannot just ignore the money supply: if it is not to lose control over the market interest rate, it has to ensure that the appropriate money supply is provided to the economy.

Note that (48) provides an easier route to the money supply that the central bank needs to provide: its growth rate has to equal the sum of the inflation rate and the output growth rate less the rate of change of velocity.

15.4.7 NK business cycle theory

The various models of the Keynesian paradigm imply that aggregate demand changes produce changes in output, as do shifts in the production function and labor supply, so that their explanations of the business cycle allow both shifts in aggregate demand and supply factors to be a potential cause of the business cycle. Therefore, from the perspective of monetary policy, monetary policy can be potentially a cause of business cycles as well as being useful in reducing their severity and duration. In fact, most central banks do habitually manipulate monetary policy to moderate inflation and the output gap, thereby subscribing to the implication of the Keynesian paradigm that monetary policy can moderate output and employment fluctuations, as well as maintain inflation within an acceptable range. This policy practice has been confirmed by the estimation of some form of the Taylor rule for many countries over many periods.

New Keynesians base their explanations of business cycles on sticky prices or sticky inflation, as opposed to the flexible price assumption of the real business cycle theory (see Chapter 14). In the NK model, the potential sources of fluctuations in output include shifts in aggregate demand due to shifts in investment, consumption, exports, fiscal and monetary policy, etc. Conversely, output fluctuations due to aggregate demand shocks can be moderated by monetary policy. Further, the new Keynesians' use of the Taylor rule on the formulation of monetary policy embeds in monetary policy an automatic stabilizer since, under this rule, an output gap (with actual output below the full-employment level) triggers an expansionary reduction in the interest rate.

Note that once both demand and supply sources of business cycle fluctuations are allowed, their relative importance is likely to vary over different business cycles and for different countries. One stance on the explanation of business cycles is the proposition that changes

47 Note that this mode of determination of the price level is shared by the monetarist, neoclassical and Keynesian macroeconomic models.

in technology, possibly due to the revolution in information technology (IT), have been the *dominant* cause of actual business cycle fluctuations in industrialized economies in recent decades, and that while aggregate demand fluctuations can cause business cycles, they have not been the dominant source. There does seem to be substantial econometric support for this proposition.

To provide one illustration out of the many available studies on this topic, Ireland (2001) tests a new Keynesian dynamic, stochastic general equilibrium model of the business cycle for the USA over the period 1959:1 to 1998:4. His model includes sticky prices (or inflation) and adopts the nominal interest rate as the monetary policy variable, with an additional assumption on the cost of adjusting physical capital. This study reports that the model performs better with costs of adjusting prices rather than inflation. He also finds that persistence in inflation is due more to exogenous real shocks, i.e. to preferences and technology, than to large adjustment costs.

15.5 Reduced-form equations for output and employment in the Keynesian and neoclassical approaches

The fundamental implication of the various forms of Keynesian models is that, in conditions of aggregate demand deficiency and less than full employment, aggregate output depends upon aggregate demand and therefore on the demand management policy variables of fiscal expenditures, taxation and the money supply. On the impact of monetary policy on real output, the Keynesian analyses imply that:

1 This impact will depend upon the existing demand deficiency in the economy, so that a linear relationship between real output and the money supply, with constant coefficients, is not a proper representation of the Keynesian implications.
2 Both the unanticipated *and* the anticipated values of the money supply – as also of the fiscal variables – will affect output equally, as against the modern classical assertion that only the unanticipated values do so.

Keynesian reduced-form equations

A simple linear equation for capturing the dependence of output on the policy variables is:

$$dy = \lambda_g(\mathrm{D}y)dg + \lambda_M(\mathrm{D}y)\mathrm{d}M \qquad \lambda_g \geq 0 \tag{49}$$

where Dy is the output gap $(y^{\mathrm{f}} - y)$, λ_g and λ_M are functional symbols, and M is the relevant monetary policy instrument. λ_M is non-negative if this instrument is the money supply and non-positive if it is the interest rate.[48] As shown in the preceding sections, λ_g and λ_M depend upon the existing demand deficiency in the economy[49] and cannot be taken to be constants.

48 In the demand-deficient cases, λ_g, λ_M will be non-zero, without being constant, while in the limiting case of long-run equilibrium (i.e. full employment), λ_g, $\lambda_M = 0$.

49 This makes them state-contingent, while the Calvo price adjustment process of the NK model makes them time-contingent since, in this process, firms adjust prices on a time schedule, though it is one which is determined by the economic environment that they face.

For a dynamic context, define $\mathrm{d}y = y_t - y_{t-1}$, $\mathrm{d}g = g_t - g_{t-1}$, $\mathrm{d}M = M_t - M_{t-1}$, so that (49) becomes:

$$y_t = [y_{t-1} - (\lambda_g(\mathrm{D}y_{t-1})g_{t-1} + \lambda_M(\mathrm{D}y_{t-1})M_{t-1})] + \lambda_g(\mathrm{D}y_{t-1})g_t + \lambda_M(\mathrm{D}y_{t-1})M_t \quad (50)$$

In line with (50), an equation popular for empirically testing the Keynesian model is:

$$y_t = a_0 + \lambda_g(\mathrm{D}y_{t-1})g_t + \lambda_M(\mathrm{D}y_{t-1})M_t + \mu_t \quad \lambda_g, \lambda_M \geq 0 \quad (51)$$

where a_0 equals $[y_{t-1} - (\lambda_g g_{t-1} + \lambda_M M_{t-1})]$, μ_t is a random term, and all variables are in logs. a_0 is sometimes replaced by a term of the form αy_{t-1} to capture persistent patterns in output, and y_t and y_{t-1} are sometimes defined to be deviations in output from its full-employment level.

Another form of the above Keynesian equation uses the deviation of unemployment from its natural rate in place of the output gap, and can be specified as:

$$y_t = a_0 + \lambda_g(u_t - u^{\mathrm{n}}{}_t)g_t + \lambda_M(u_t - u^{\mathrm{n}}{}_t)M_t + \mu_t \quad (52)$$

where u is the unemployment rate, u^{n} is the natural (or full-employment) rate of unemployment and λ_g and λ_m are functional symbols. $\lambda_g(u_t - u^{\mathrm{n}}{}_t)$ and $\lambda_M(u_t - u^{\mathrm{n}}{}_t)$ need not be linear functions.

Comparison of the NK and modern classical estimating equations

Note that the modern classical equation corresponding to (52) is:

$$y_t = y^{\mathrm{LR}} + b_g(g_t - g^{\mathrm{e}}{}_t) + b_M(M_t - M^{\mathrm{e}}{}_t) + \mu_t \quad (53)$$

where $b_g, b_M > 0$ and the superscript e indicates the expected value of the variable in question. In expectational equilibrium with $g_t = g^{\mathrm{e}}{}_t$ and $M_t = M^{\mathrm{e}}{}_t$, (53) becomes:

$$y_t = y^{\mathrm{LR}} + \mu_t \quad (54)$$

which states the modern classical conclusion that if there are no errors in expectations, the only deviations around the full-employment output that can occur have to be random ones.

In the modern classical model, since any impact of money supply increases and fiscal deficits on output must *first* cause an increase in individual prices or the price level, the more appropriate equations for testing the modern classical model are:

$$y_t = a_0 + \gamma_1(P - EP_t) + \gamma_P EP_t + \mu_t \quad (55)$$

$$y_t = a_0 + \gamma_1(\pi - E\pi_t) + \gamma_\pi E\pi_t + \mu_t \quad (56)$$

where EP and $E\pi$ are the rationally expected values. The equilibrium version of the modern classical model implies that the estimated values of γ_P and γ_π will be zero, so that monetary policy changes will be neutral.

15.6 Empirical validity of the new Keynesian ideas

One way to judge the empirical validity of the NK model is to compare its implications with the stylized facts listed in Chapters 1 and 14 on the relationship between money, inflation and output. For the long run, the NK model holds that there is no relationship between money or inflation and output, as does the modern classical school, so that the stylized short-run facts are really the relevant ones for judging their relative validity. For the short run, the NK model clearly explains more of the stylized facts or does so better than the modern classical model. In particular, it explains the hump-shaped pattern of the impact of monetary changes on output and the longer lag in the impact on money of inflation than on output (Nelson, 1998; Sims, 1992; Christiano *et al*., 1999). Wong (2000) finds that long-run monetary neutrality and short-run price stickiness hold for the United States. While the NK approach implies an unchanging pattern of impulse responses over time to monetary shocks, Wong finds that the actual responses differ during different episodes and are stronger for negative shocks than for positive ones.

Rudd and Whelan (2003) test the forward-looking NK output equation, which asserts that the current inflation is positively related to the future output gap. They find that this equation does very poorly for US data: empirically, current inflation is negatively, not positively, related to the future output gap. One reason for this poor performance is that there is a high degree of persistence in inflation, which depends heavily on its own lagged values (see also Maria-Dolores and Vazquez, 2006).

The NK school relies on sticky prices and nominal wages for the dynamic impact of monetary policy on output. Christiano *et al*. (2001) find that wage stickiness rather than price stickiness seems to be the more relevant factor in explaining this dynamic impact. However, Mankiw (2001, p. C52) points out that the NK price adjustment equation based on sticky prices "is completely at odds with the facts. In particular, it cannot come even close to explaining the dynamic effects of monetary policy on inflation and unemployment." Mankiw lists three invalid implications of the NKPC. One, it implies that a fully credible disinflation causes an increase in output, since its announcement leads firms to reduce their price increases even before the money supply growth rate is reduced. This causes an increase in the real money supply, which leads to higher output and lower unemployment. This is invalid since inflation rises in booms and falls in recessions and the commonly observed result of a decrease in the money supply is disinflation accompanied by a rise in unemployment. Two, in the NK price adjustment process, individual prices adjust intermittently so that the price level adjusts slowly to shocks. However, the change in prices as measured by the inflation rate adjusts instantly, so that the model does not generate persistence in inflation. What is observed in reality is that the impact of a shock to inflation builds up gradually over several quarters, so there is considerable persistence in inflation. Three, the NKPC does not generate plausible impulse response functions in inflation and unemployment following a monetary policy shock. Empirical data shows that a monetary policy shock affects unemployment for some time after the shock, while the shock has a *delayed and gradual* effect on inflation. Mankiw concludes that a backward-looking price adjustment process, with inflation inertia, is needed to explain the observed behavior of inflation.

Many empirical studies find support for some form of the Taylor rule. However, Levin *et al*. (1999, 2001) report that their estimates from US data for five macroeconomic models show that a simple version of the inflation and output-targeting rule for the US economy does quite well. Further, a simple autoregressive model of the interest rate has sometimes performed better than the Taylor rule, as Depalo (2006) reports for Japan. Moreover, estimates of the

Taylor rule tend to show that its coefficients do shift with changes in the leadership of the central bank. This is quite plausible since these coefficients reflect central bank preferences on responses to inflation and the output gap, etc., as the NK derivation of the Taylor rule shows.

Note that, to derive their particular form of the Taylor rule, NK models assume that the true forms of the central bank's objective function and the model are known. These are rarely, if ever, known, so that the superior performance of a specific optimal Taylor rule, derived from a specific model and an objective function, relative to simple Taylor rules, cannot be taken for granted. Given the uncertainty about the future course of the economy and errors in forecasting, it may be better to just specify the central bank's objectives on inflation, output and any other targets, while leaving their actual coefficients to vary with the central bank's appraisal of the circumstances. In this case, the central bank would commit itself only to pursuing certain objectives rather than to a set instrument rule such as a specific Taylor rule. Svensson (2003) provides an excellent and detailed discussion of this issue. This conclusion is consistent with that in Chapter 12 on time consistency and credibility: the conclusion there was that, given uncertainty about the appropriate form of the economy's constraints and how they might evolve over time, it might be preferable for the central bank to commit itself to objectives in the context of intertemporal reoptimization rather than to a precise time-consistent path of policies or to a policy rule.

15.7 Conclusions

This chapter has presented three versions of the Keynesian model and a neoKeynesian model. These attest to its evolutionary nature. The future is likely to see additional contributions and versions of the general Keynesian stance that money is not neutral in the short run, while it could – and is likely to be – neutral in the long run.

We arrange the concluding comments for this chapter into various categories, as follows.

Keynes on the wage bargain and the rigidity of wages

A major point of disagreement between the classical and Keynesian ideas is on the nature of the labor market. Keynes himself considered his objections to the classical model's assumptions on the labor market as being the most fundamental departures from classical ideas. He expressed his ideas on this as follows:

> But the … more fundamental objection [to the classical model] … flows from our disputing the assumption that the general level of real wages is directly determined by the character of the wage bargain. … For there may be no method available to labor as a whole whereby it can bring the wage-goods equivalent of the general level of money-wages into conformity with the marginal disutility of the current volume of employment. There may exist no expedient by which labor as a whole can reduce its real wage to a given figure by making revised money bargains with the entrepreneurs. …
>
> Though the struggle over money-wages between individuals and groups is often believed to determine the general level of real wages, it is, in fact, concerned with a different object. Since there is imperfect mobility of labor, and wages do not tend to an exact equality of net advantage in different occupations, any individual or group of individuals, who consent to a reduction of money-wages relatively to others, will suffer a relative reduction in real wages, which is a sufficient justification for them to resist it. On the other hand it would be impracticable to resist every reduction of real wages,

> due to a change in the purchasing-power of money which affects all workers alike; and in fact reductions of real wages arising in this way are not, as a rule, resisted unless they proceed to an extreme degree.
>
> (Keynes, 1936, pp. 12–15).

On the rigidity of nominal or real wages

The Keynesian nominal wage model, which assumes that nominal wages are rigid or that the labor supply depends on nominal rather than real wages, was popular as *the* Keynesian model in the early 1940s and 1950s. However, Leijonhufvud (1967) showed that Keynes did not make such an assumption, and that, for Keynes, in conditions of excessive unemployment, reducing real wages through induced inflation was preferable to nominal wage reductions (which were resented by workers and caused industrial unrest), so that the avoidance of nominal wage declines was a policy recommendation. It is now generally accepted: Keynes did not assume that nominal wages were rigid, either as an a priori assumption or because of the nature of the nominal wage bargain between firms and workers. Further, Keynes did not assume that workers based their supply behavior on nominal rather than real wages and thereby suffered money illusion.

The distinctive nature of labor markets was a fundamental part of Keynes's ideas. One interpretation of those ideas led to the Phillips curve, as illustrated by the following quotations from James Tobin (1972), an eminent economist in the Keynesian tradition of the 1960s to the 1980s.

Tobin on wages and unemployment

> Unemployment is, in this model as in Keynes reinterpreted, a disequilibrium phenomenon. Money wages do not adjust rapidly enough to clear all labor markets every day.
>
> The overall balance of vacancies and unemployment is determined by aggregate demand, and is therefore in principle subject to control by overall monetary and fiscal policy. Higher aggregate demand means fewer excess supply markets and more excess demand markets, accordingly less unemployment and more vacancies.
>
> (Tobin, 1972, pp. 9–10).

The critical role of dynamic analysis when aggregate demand falls

The introduction of dynamic analysis did away with the Keynesians' need to assume the rigidity of prices and nominal wages. For such analysis, given a fall in aggregate demand, the central issue is the nature of the individual firm's response to a fall in the demand for its product and the nature of the response of the worker who is laid off or whose job no longer seems to be secure, in a context where the numerous markets of the economy cannot realistically be assumed to come into macroeconomic equilibrium instantly. This is a shift in the debate from comparative static to dynamic analysis. There can be numerous plausible dynamic paths corresponding to any comparative static macroeconomic model, and not all necessarily lead to full employment or do so instantly. This implies a role for Keynesian demand management policies, depending upon the state of the economy and the speed at which it is expected to redress deficient demand or involuntary unemployment.

We illustrate the central issues from this perspective by the following two quotes.

Leijonhufvud on Keynes's methodology

> Keynes's theory was dynamic. His model was static. The method of trying to analyze dynamic processes with a comparative static analysis apparatus Keynes borrowed from Marshall. … The initial response to a decline in demand is a quantity adjustment. … The strong assumption of "rigid" wages is not necessary to the explanation of such system behavior. *It is sufficient only to give up the equally strong assumption of instantaneous price adjustments*.
>
> (Leijonhufvud, 1967, pp. 401–03; italics added).

Patinkin on effective demand analysis

> Involuntary unemployment can have no meaning within the confines of static equilibrium analysis. Conversely, the essence of dynamic analysis is *involuntariness*: its domain consists only of positions off the [notional] demand or supply curves. …
>
> First, we see that involuntary unemployment can exist in a system of perfect competition and wage and price flexibility. … Second, we see that a deficiency in commodity demand can generate a decrease in labor input without requiring a prior increase in the real wage rate.
>
> (Patinkin, 1965, pp. 323–24).

> And the assumption … that, granted flexibility, these [dynamic] forces will restore the economy to a state of full employment, is an assumption that … [a full employment] equilibrium position always exists and that the economy will always converge to it. More specifically, it is an assumption that just as the "market" can solve the system of excess demand equations (of the neoclassical model), when the level of real income is held constant during the *tâtonnement*, so can it solve it when the level of real income (and hence employment) is also permitted to vary.
>
> (Patinkin, 1965, p. 328).

The points made in these quotes are now generally accepted. Keynes's analysis was not comparative static or Walrasian equilibrium analysis based on the assumption of instantaneous market-clearing *price adjustments with perfect competition* in response to excess demand or supply, but focused on the *dynamic adjustments* made by firms when these conditions did not hold. If the Walrasian market adjustments in prices towards equilibrium were slow, then, quite plausibly, a shortfall in demand would result in reductions in output and employment for periods of significant length. If the economy was beset by fresh disturbances arising frequently, such as through bouts of pessimism or optimism about the future among firms' managers and households, the disequilibrium state would be a persistent phenomenon, with varying levels of employment or output.

NeoKeynesian and new Keynesian economics

NeoKeynesian economics came into being in an attempt to rebuild the Keynesian framework after the decline of faith in the 1970s in the Keynesian models and their policy prescriptions, and the resurgence of classical economics in the 1980s and 1990s. NeoKeynesian economics

was not really an integrated macroeconomic model, but rather a collection of ideas and theories. One of these was the efficiency wage hypothesis, which asserted the short-run rigidity of real wages, in contrast to that of nominal wages. Another neoKeynesian theory was the menu cost theory, which provided a new basis for the short-run rigidity of prices through its hypothesis of menu costs in monopolistic competition. Its other theories included the implicit contract theory based on long-term contracts in a context of firm-specific labor skills, implying labor hoarding and variations in work effort over the business cycle.

New Keynesians seek to provide an integrated macroeconomic framework, which has adopted as its major component the staggered slow price adjustment, due to menu costs, made by monopolistically competitive firms. It ignores or downplays the efficiency wage theory and the implicit contract theory, which had been part of neoKeynesian economics. A second component of the new Keynesian model is its adoption of a Taylor rule for the pursuit of monetary policy. However, the essential distinctiveness of the new Keynesian model from neoKeynesian economics lies in its methodology, which derives the various equations of the model from microeconomic foundations with stochastic intertemporal optimization, rational expectations and general equilibrium. This methodology assigns greater weight to the impact on current values of the future values of the variables, while assigning much less weight to the impact of the past values of the variables. In doing so, new Keynesians have discarded critical parts of the earlier Keynesian models, such as labor market behavior, deficient demand analysis and involuntary unemployment, as well as the possibility of the failure of markets to clear. The new Keynesian model is a relatively new one and its empirical validity is still very much in dispute.

Mankiw (2001) summarizes the evidence on the relationship between inflation and unemployment in the following:

> Almost all economists today agree that monetary policy influences unemployment, at least temporarily, and determines inflation, at least in the short run. … Price stickiness can explain why society faces a short-run tradeoff between inflation and unemployment.
>
> The bad news is that the dynamic relationship between inflation and unemployment remains a mystery. The so-called "new Keynesian Phillips curve" is appealing from a theoretical standpoint, but it is ultimately a failure. It is not at all consistent with the standard stylized facts about the dynamic effects of monetary policy, according to which monetary shocks have a delayed and gradual effect on inflation. We can explain these facts with traditional backward-looking models of inflation–unemployment dynamics, but these models lack any foundation in the microeconomics theories of price adjustment.
>
> (Mankiw, 2001, p. C59).

And finally, what are we to believe?

The classical and Keynesian schools represent different views of the dynamic nature of the capitalist economy. The former views it as being in full-employment equilibrium or close to it, with the dynamic forces providing a strong tendency to return to full employment after any deviation. Keynesian schools allow the possible existence of full employment but are concerned that the economy does not always, or most of the time, perform at full employment. Historically, faith in these positions has tended to vary considerably. The Great Depression of the 1930s in industrial economies destroyed faith in the classical and neoclassical belief in a self-regulating economy The fairly stable macroeconomic performance of such economies in the 1950s and 1960s, though with active Keynesian demand management policies, produced

shallow and short-lived recessions and led to a slow revival first of classical economics under the rubric of the Keynesian–neoclassical synthesis. The Keynesian policy errors in the 1970s, resulting in stagflation, tended to restore faith in the general classical position. While the dominant school in the last three decades of the twentieth century seemed to be the classical one, the Keynesian doctrines, rejuvenated and reformulated, seemed to reclaim dominance toward the end of the twentieth century. Its currently popular, or rather fashionable, form is that of the NK model, which adopts the market-clearance, general equilibrium, rational expectations agenda of the modern classical school, but adds to it market imperfections and sticky prices or information.

The economics profession does not possess empirical evidence sufficient to convince all economists to accept one paradigm and discard the other, and the performance of the economy seems to suit one paradigm at one time and the other one at other times, so that many economists keep an open mind, applying their overall knowledge depending upon the state of the economy. Most economists also maintain a fair degree of skepticism.

Given the uncertainty about the true nature of the macroeconomy and disagreements among economists about the correct workhorse for the economy, the practitioners of monetary policy, the central banks, do not follow a time-invariant commitment strategy to any particular model or its implications. Rather, as Chapter 12 argued, they follow a commitment strategy only with regard to their objectives. In their policies, they follow an active agenda, as the Keynesians in general recommend, for guiding the short-run evolution of output and employment through changes in the money supply and/or interest rates, while subscribing to the classical economists' belief that monetary policy cannot affect the long-run evolution of the real variables.

The current fashion in macroeconomics is to base its foundations strictly and solely on microeconomic intertemporal stochastic general equilibrium foundations. The following quote provides a caution against this practice.

> Economists often aspire to make our discipline like physics. Just as there are today two "economicses" – micro and macro … – there are also two "physicses": quantum theory, which describes the behavior of the tiniest particles of matter, and Newtonian mechanics (as amended by the theory of relativity), which applies to larger bodies. One of the challenges that physicists face is to integrate the two. As the distinguished mathematician Roger Penrose has pointed out, however, the way to do it is clearly not simply to take the principles of quantum theory and apply them wholesale to larger bodies. Doing so leads, in Penrose's classic example, to concluding that a basketball can be in two places at once. … Simply applying to aggregate economies what we know about the behavior of rational, profit- or utility-maximizing individual agents leads to patent contradictions of the economic world in which we live.
>
> (B.M. Friedman, 2003, p. 10).

Summary of critical conclusions

- An abiding theme of the Keynesian paradigm, originating with Keynes's *The General Theory*, is the failure of the economy to attain Walrasian general equilibrium. Many of its models assert that this failure is especially symptomatic of the labor market, so that involuntary unemployment is a common occurrence in real-world economies.

- ❖ Early (1940s and 1950s) Keynesian models were based on nominal wage rigidity or price illusion by labor.
- ❖ In the 1960s and 1970s, Keynesian models were often based on the Phillips curve.
- ❖ The Keynesian effective demand model posits that the rational dynamic responses by firms and households to conditions of inadequate demand and involuntary unemployment do not always take the economy to full employment or do so within an acceptable period.
- ❖ The neoKeynesian theories rely on rational long-run behavior, resulting in implicit contracts, staggered wage contracts, sticky prices, menu costs, etc.
- ❖ The new Keynesians base their macroeconomic model on the microeconomic foundation of forward-looking, optimizing economic agents holding rational expectations and with the economy operating in general equilibrium. Their distinctiveness from the modern classical model, which has a similar methodology, lies in that they assume monopolistically competitive firms, sticky prices and the Taylor rule for the central bank's monetary policy.

Review and discussion questions[50]

1. "In response to demand shocks, short-term quantity adjustments occur earlier than price adjustment at the level of both the firm and the economy." Discuss the relevant theory behind this statement. Also, discuss its empirical validity at the macroeconomic level.
2. Discuss in the context of the effective demand and Phillips curve Keynesian models: excluding dynamic effects, an increase in the stock of money and a fall in nominal wages have essentially the same effects at a time of involuntary unemployment.
3. Discuss in the context of the neo- and new Keynesian models: excluding dynamic effects, an increase in the stock of money and a fall in nominal wages have essentially the same effects at a time of involuntary unemployment.
4. (a) Describe a simple fixed-price short-run macroeconomic model (with flexible nominal wages) and compare it with a conventional market-clearing model. Compare their predictions for the effectiveness of monetary and fiscal policies.
   (b) Describe a simple short-run macroeconomic model with flexible prices but fixed nominal wages and compare it with a conventional market-clearing model (with flexible nominal wages). Compare their predictions for the effectiveness of monetary and fiscal policies.
   (c) Describe a simple short-run macroeconomic model with a fixed price and fixed nominal wages and compare it with a conventional market-clearing model. Compare their predictions for the effectiveness of monetary and fiscal policies.
5. You are given the following fixed-price, closed-economy, IS–LM model:

   $$IS: \quad y = c[(1 - t_1)(y + b/r), M + b/r] + i(r, y) + g$$

   $$LM: \; M = m^{\mathrm{d}}(R, y, M + b/r)$$

   Fisher equation: with an exogenously specified expected inflation rate at zero.

50 Unless a symbol is specifically defined in a question, its definition is as in the text.

The government's budget constraint is:

$$dM + db/r = g - t_1(y + b/r)$$

where b is the number of consols, each paying \$1 per period in perpetuity. P is normalized to unity. Wealth is held only in money and bonds.

(a) Explain the differences between the IS and LM relationships in this question and those used in this chapter and Chapter 13.
(b) Explain the government budget constraint.
(c) Using IS–LM diagrams, derive the short-run and long-run equilibrium effects on output of a permanent increase in g financed by (i) money creation, (ii) bond creation. Under what conditions are these policies stable?
(d) How are your results affected if bonds are not part of net wealth?
(e) Does this model explain some of the differences between the Monetarists and the Keynesians on the relative efficacy of monetary versus fiscal policies?

6. Suppose that business pessimism reduces investment such that aggregate demand becomes less than full employment income at all *non-negative* rates of interest. Use IS–LM analysis to answer the following:

(a) Are there positive equilibrium levels of y, r and P in the neoclassical model?
(b) Are there positive equilibrium levels of y, r and P in the Keynesian fixed-price model? In the Keynesian nominal wage model (without fixed prices)?

What processes will take the economy to these levels?

7. "From the time of Say and Ricardo the classical economists have taught us that the supply creates its own demand … (and) that an individual act of abstaining from consumption necessarily leads to … the commodities thus released … to be invested … so that an act of individual saving inevitably leads to a parallel act of investment … Those who think this way are deceived. They are fallaciously supposing that there is a nexus which unites decisions to abstain from consumption with decisions to provide for future consumption, whereas the motives which determine the latter are not linked with the motives which determine the former." (Keynes, 1936, pp. 18–21). Explain this statement.

If investment and saving depend on different determinants, what are these determinants and how is the equality of saving and investment in the economy ensured? Or does it also become an identity? If it is not an identity, outline the possible scenario of the likely adjustment pattern in the economy following an exogenous decrease in consumption.

8. Suppose the central bank pegs the price level by using money supply changes through open market operations. Present the IS–LM analysis incorporating this money supply rule and show the implications for the money supply, aggregate demand and output of an exogenous increase in autonomous consumption. Is the effect on interest rates less or greater under this money supply rule than if the money supply were held constant.
9. Start with the neoclassical model and assume that its equilibrium solution is (y^f, n^f, r^*, P^*). Suppose a reduction in investment reduces aggregate demand. Discuss the following:

(a) Within the context of the *neoclassical* model, analyze the behavior of firms if they face imperfect competition and are hit with a fall in the demand for their products. If this analysis shows that employment is reduced below n^f, present the likely

consumption response of households. If these responses of firms and households imply a movement away from (y^f, n^f, r^*, P^*), what equilibrating mechanisms will come into play to bring the economy back to (y^f, n^f, r^*, P^*)? Which do you think is more powerful and has a faster response: the economy's equilibrating mechanisms or the (contrary) responses of firms and households which worked to take the economy away from (y^f, n^f, r^*, P^*)?

(b) In the context of a *Keynesian* model with nominal wage contracting, redo the questions in (a).

(c) Within the context of a neoKeynesian model, redo the questions in (a).

10. "The *classical theory* dominates the economic thought, both practical and theoretical, of the governing and academic classes of this generation, as it has for a hundred years past. … [But it is] applicable to a special case only and not to a general case, the situation it assumes being a limiting point of the possible positions of equilibrium." (Keynes, 1936, p. 3).

    (a) What are the traditional classical (also neoclassical and modern classical) and Keynesian definitions of equilibrium? How are they related? Can there exist an underemployment equilibrium in their models under each of these definitions?

    (b) If you adopt the Keynesian definition of equilibrium, was the traditional classical model a special case of any of the Keynesian models? Of the new Keynesian model?

    (c) Are the neoclassical and modern classical models also special cases of any of the Keynesian models?

11. Keynes argued that an economy could be in equilibrium with a substantial amount of involuntary unemployment, but other economists took the stand that an equilibrium in which an important market does not clear is a contradiction in terms. Explain the notions of equilibrium involved, Keynes's justification for his position, and his opponents' justification for theirs.

12. Distinguish between Keynesian unemployment caused by an aggregate demand deficiency and classical unemployment due to real wages being above the full-employment level. What can monetary policy do to reduce each of these?

13. One way of capturing the degree of indexation of nominal wages is by specifying the wage contract as:

$$W - W_0 = \alpha(P - P_0) \qquad 0 < \alpha < 1$$

    where W_0 and P_0 are the nominal wage and price levels at the time of the negotiation of the wage contract and all variables are in logs. $\alpha = 1$ indicates full indexation.

    Compare the aggregate supply curve when W is indexed to P with $\alpha < 1$ with the curves in the fixed nominal wage Keynesian model and the flexible nominal wage neoclassical model. What are the implications of $\alpha = 1$ for the responsiveness of output and the price level to (a) aggregate demand shocks, (b) aggregate supply shocks? In this case, would the aggregate supply curve differ from the aggregate supply curve with a flexible nominal wage?

    Show that real output is less sensitive and the price level more sensitive to changes in the money supply if α is greater.

14. J. R. Hicks, in the 1937 article in which he proposed the IS–LM analysis, argued that Keynes's *General Theory* did not represent a major break with the classical tradition.

In particular, he maintained that the main insight it contained was into the conditions existing during a depression or a deep recession. Was this claim valid?

Have Keynesians contributed anything further since Keynes's *General Theory*? Does the above claim apply to the various Keynesian models?

Does the modern classical approach provide an adequate analysis of the economic conditions in recessions and depressions, or does the profession still need the Keynesian approaches for its analysis?

15. "Keynes argued that wage stickiness was probably a good thing, that wage and price flexibility could easily be destructive of real economic stability. His reasoning went like this. In a monetary economy, the nominal interest rate cannot be negative. Hence the real interest rate must be at least equal to the rate of deflation. … If wages and prices were to fall freely after a contractionary shock, the real interest rate would become very large at just the wrong time, with adverse effects on investment. The induced secondary contraction would only worsen the situation." (Solow, 1980). Discuss the validity of these arguments. Do they apply also to the modern classical model?
16. Keynes (1936) argued that, from a policy perspective, everything that can be achieved by a nominal wage cut can be more effectively achieved through an appropriate monetary policy.

    (a) Does this statement hold in the deficient-demand Keynesian model for a negative shock to (i) aggregate demand and (ii) aggregate labor productivity?
    (b) Does this statement hold in the new Keynesian model for a negative shock to (i) aggregate demand and (ii) aggregate labor productivity?

17. "Keynesianism and new Keynesianism are fundamentally inconsistent in so far as their wage hypotheses are concerned. Keynesianism asserts nominal wage rigidity, at least downwards, while the new Keynesianism does not allow either nominal or real wage rigidity." Discuss.
18. In the last two decades of the twentieth century, many economists believed that Keynesian economics give little or even wrong prescriptions for dealing with the current economic problems in the United States (or British or Canadian) economy. What justifies such comments? What are your views on this issue and how would you justify them?
19. How does the new Keynesian model differ from the earlier Keynesian deficient-demand model? How does it differ from the modern classical one? Which of the three models would explain involuntary unemployment in a recession following a fall in aggregate demand?
20. "Every major inflation has been produced by monetary expansion" (Friedman, 1968). Does this assertion hold for the NK model, whose equations do not even include a monetary aggregate as a variable?
21. "Because monetary shocks have a delayed and gradual impact on inflation, in essence we experience a credible announced disinflation every time we get a contractionary shock. Yet we don't get the boom that the (new Keynesian) model says should accompany it. This means that something is fundamentally wrong with this model." Discuss the validity of the preceding observation. In particular, is this observation valid for contractions in monetary policy that were unanticipated when the policy was first announced? What does the new Keynesian model imply on this issue? If it is fundamentally wrong, how might the new Keynesian model be modified to fit the facts?

22. "It is now well established that a contractionary monetary shock increases unemployment before reducing inflation and that the peak impact on unemployment precedes the peak impact on inflation." Discuss how well these observations are explained by (i) the modern classical model, (b) the new Keynesian model, and (c) any one other Keynesian model of your choice.

References

Ball, L., Mankiw, N.G. and Romer, D. "The new Keynesian economics and the output–inflation trade-off." *Brookings Papers on Economic Activity*, 19, 1988, pp. 1–65.

Blanchard, O. "What do we know about macroeconomics that Fisher and Wicksell did not?" *Quarterly Journal of Economics*, 65, 2000, pp. 1375–409.

Boschen, J.F. and Weise, C.L. "What starts inflation: evidence from the OECD countries." *Journal of Money, Credit and Banking*, 35, 2003, pp. 323–49.

Calvo, G. "Staggered prices in a utility maximizing framework." *Journal of Monetary Economics*, 12, 1983, pp. 383–98.

Christiano, L.J., Eichenbaum, M. and Evans, C. "Monetary policy shocks: what have we learned and to what end?" In J. Taylor and M. Woodford, eds, *Handbook of Macroeconomics*, Vol. 1A. Amsterdam: Elsevier North-Holland, 1999, pp. 65–148.

Christiano, L.J., Eichenbaum, M. and Evans, C. "Nominal rigidities and the dynamic effects of a shock to monetary policy." *NBER Working Paper no.* 8403, 2001.

Clarida, R., Gali, J. and Gertler, M. "The science of monetary policy: a new Keynesian perspective." *Journal of Economic Literature*, 37, 1999, pp. 1661–707.

Clarida, R., Gali, J. and Gertler, M. "Monetary policy rules and macroeconomic stability: evidence and some theory." *Quarterly Journal of Economics*, 115, 2000, pp. 147–80.

Clower, R. "The Keynesian counter-revolution: a theoretical appraisal." In F.H. Hahn and F.P.R. Brechling, eds, *The Theory of Interest Rates*. London: Macmillan, 1965.

Depalo, D. "Japan: the case for a Taylor rule? A simple approach." *Pacific Economic Review*, 11, 2006, pp. 527–46.

Eichenbaum, M. and Fisher, J.D.M. "Estimating the frequency of price re-optimization in Calvo-style models." *Journal of Monetary Economics*, 54, 2007, pp. 2032–47.

Friedman, B.M. "The LM curve: a not-so-fond farewell." *NBER Working Paper* no. 10123, 2003.

Friedman, M. "The role of monetary policy." *American Economic Review*, 58, 1968, pp. 1–17. Reprinted in Milton Friedman, *The Optimum Quantity of Money and Other Essays*. Chicago: Aldine Publishing Co. 1969, p. 106.

Gali, J. "New perspectives on monetary policy, inflation and the business cycle." *NBER Working Paper* no. 8767, 2002.

Galí, J. and Gertler, M. "Inflation dynamics: a structural econometric analysis." *Journal of Monetary Economics,* 44, 1999, pp. 195–222.

Hafer, R.W., Haslag, J.H. and Jones, G. "On money and output: Is money redundant?" *Journal of Monetary Economics*, 54, 2007, pp. 945–54.

Handa, J. *Monetary Economics*. London, Routledge, 2000.

Hicks, J.R. "Mr. Keynes and the classics: a suggested interpretation." *Econometrica*, 5, 1937, pp. 147–59.

Ireland, P.N. "Sticky-price models of the business cycle: specification and stability." *Journal of Monetary Economics*, 47, 2001, pp. 3–18.

Keynes, J.M. *A Tract on Monetary Reform*. London: Macmillan, 1923.

Keynes, J.M. *The General Theory of Employment, Interest and Money*. New York: Macmillan, 1936.

Leijonhufvud, A. "Keynes and the Keynesians." *American Economic Review, Papers and Proceedings*, 57, May 1967, pp. 401–10.

Leijonhufvud, A. *On Keynesian Economics and the Economics of Keynes*. New York: Oxford University Press, 1968.

Levin, A., Wieland, V. and Williams, J.C. "Robustness of simple monetary policy rules under model uncertainty." In J.B. Taylor, ed., *Monetary Policy Rules*. Chicago: University of Chicago Press, 1999, pp. 263–99.

Levin, A., Wieland, V. and Williams, J.C. "The performance of forecast-based monetary policy rules under model uncertainty." Working Paper 2001-39, *Board of Governors of the Federal Reserve System*, 2001.

Mankiw, N.G. "The inexorable and mysterious tradeoff between inflation and unemployment." *Economic Journal*, 111, 2001, pp. C45–C61.

Mankiw, N.G. "Pervasive stickiness." *American Economic Review*, 96, 2006a, pp. 164–9.

Mankiw, N.G. "Sticky information in general equilibrium." *NBER Working Paper* no. 12605, 2006b.

Mankiw, N.G. and Reis, R. "Sticky information versus sticky prices: a proposal to replace the new Keynesian Phillips curve. *Quarterly Journal of Economics*, 117, 2002, pp. 1295–328.

Maria-Dolores, R. and Vazquez, J. "How does the new Keynesian monetary model fit in the U.S. and the Eurozone? An indirect inference approach." *Topics in Macroeconomics*, 6, 2006, article 9, pp. 1–49.

Nelson, E. "Sluggish inflation and optimising models of the business cycle." *Journal of Monetary Economics*, 42, 1998, pp. 302–22.

Nelson, E. "Direct effects of base money on aggregate demand: theory and evidence." *Journal of Monetary Economics*, 49, 2002, pp. 687–708.

Okun, A. *Prices and Quantities: A Macroeconomic Analysis*. Washington DC: Brookings Institution, 1981.

Patinkin, D. *Money, Interest and Prices*. New York: Harper and Row, 1965.

Phillips, A.W. "The relation between unemployment and the rate of change of the money wage rates in the U.K. 1861–1957." *Economica*, 25, 1958, pp. 283–99.

Rotemberg, J. and Woodford, M. "Interest rate rules in an estimated sticky price model." In J.B. Taylor, ed., *Monetary Policy Rules*. Chicago: University of Chicago Press, 1999.

Rudd, J. and Whelan, K. "Can rational expectations sticky price models explain inflation dynamics." at www.federalreserve.gov/pubs/feds/2003/200346, 2003.

Rudebusch, G.D. "Federal reserve interest rate targeting, rational expectations and the term structure." *Journal of Monetary Economics,* 35, 1995, pp. 245–74.

Rudebusch, G.D. and Svensson, L.E.O. "Eurosystem monetary targeting: lessons from US data." *European Economic Review*, 46, 2002, pp. 417–42.

Shapiro, C. and Stiglitz, J.E. "Equilibrium unemployment as a worker discipline device." *American Economic Review*, 74, 1984, pp. 433–44.

Sims, C.A. "Interpreting the time series facts: the effects of monetary policy." *European Economic Review*, 36, 1992, pp. 975–1000.

Solow, R. "On theories of unemployment." *American Economic Review*, 70, 1980, pp. 1–11.

Svensson, L.E.O. "What is wrong with Taylor rules? Using judgment in monetary policy through targeting rules." *Journal of Economic Literature*, 41, 2003, pp. 426–77.

Taylor, J.B. "Discretion versus policy rules in practise." *Carnegie-Rochester Conference Series on Public Policy*, 39, 1993, pp. 195–215.

Tobin, J. "Inflation and unemployment." *American Economic Review*, 62, 1972, pp. 1–18.

Walsh, C. *Monetary Theory and Policy*, 2nd edn. Cambridge, MA: MIT Press, 2003.

Wong, K. "Variability in the effects of monetary policy on economic policy." *Journal of Money, Credit, and Banking*, 32, 2000, pp. 179–98.

Woodford, M. "How important is money in the conduct of monetary policy?" *NBER Working Paper* no. 13325, 2007.

Yellen, J.L. "Efficiency wage models of unemployment." *American Economic Review*, 74, 1984, pp. 200–5.

16 Money, bonds and credit in macro modeling

For the short-run macroeconomic analysis, this chapter differentiates between two types of non-monetary financial assets: bonds and credit, of which bonds are long-term financial instruments and credit represents short-term assets. The distinctive aspect of credit relies upon adverse selection, moral hazard, and monitoring and agency costs, which provide a basis for credit rationing in quantity. Credit is treated in this chapter as the variable element of working capital in the short-run. The distinctive impact of credit on economic activity is designated the credit channel.

An important element of credit is bank loans. The distinctive impact of bank loans on economic activity occurs through the bank lending channel.

Key concepts introduced in this chapter

- Adverse selection
- Moral hazard
- Monitoring and agency costs
- Credit versus bonds
- Credit rationing
- Credit market equilibrium
- Bank loans
- Working capital
- Indirect production function

The IS–LM and IS–IRT models belong to a context that assumes perfect substitution among non-monetary financial assets, which means that all non-monetary assets, as well as their different types, are perfect substitutes, so that the distinctive elements of bonds, stocks and loans, as well as the distinctions between short- and long-term bonds, government and corporate bonds, etc., can be omitted from the analysis, and all such assets can be encompassed in a composite financial asset which is labeled "bonds." Hence, the assumption of perfect financial markets leaves the macroeconomic model with only two financial assets, money (which functions as a medium of payments) and bonds (which do not), and only two financial

markets, which are the markets for money and bonds.[1] In the case where the central bank sets the money supply, the money market is specified by the LM equation (see Chapter 13). In the standard IS–LM analysis, an increase in the money supply decreases the interest rate on bonds, which increases investment and aggregate demand.

The extension of the perfect markets (i.e. perfect competition, perfect information and market efficiency) hypothesis to non-monetary financial assets implies the irrelevance of the composition of financial assets in firms' and households' portfolios and liabilities. For such a scenario, Modigliani and Miller (1958) provided the Modigliani–Miller theorem, which showed that, under perfect markets and ignoring differential tax treatments, the firm's combination of bond and equity financing was irrelevant to its output and employment, as well as to its profits which would depend on technology, inputs and consumer tastes. Further, Fama (1980) showed that whether the public holds money, bonds or stocks is irrelevant to real economic outcomes, which depend only on technology, tastes and resources. Hence, under the perfect financial markets hypothesis, financial assets and their distinctive characteristics were irrelevant for the real variables of the economy and there was no need to separate them into money, bonds, stocks, loans, etc. An intuitively unrealistic implication of these results is that credit crunches[2] and bank panics/runs[3] would have no impact on output and employment, even though such impact is often observed in recessions.[4]

As against the assumption of perfect markets, the central theme of neoKeynesian and new Keynesian economics is market imperfections.[5] Applied to financial markets, market imperfections between bonds, stocks and loans imply that they are not perfect substitutes for each other since each has quite distinctive characteristics. Further, each of these assets has many distinctive sub-categories. For example, among bonds there are short-term and long-term bonds, and there are low-risk bonds issued by some governments and high-risk ones issued by some corporations. Therefore, an extreme emphasis on market imperfections would lead to the consideration of very many different non-monetary financial assets.

However, aggregation is essential to macroeconomics, so that it needs to use the smallest number of composite goods that will adequately explain the desired macroeconomic aspects of the economy. For financial assets, the general consensus until about the 1980s was that two composite assets, money and bonds, are adequate for explaining the impact of monetary

1 In the limiting case, the assumption of perfect markets, combined with perfectly competitive and efficient markets, implies the neutrality of both money and bonds, so that all financial assets become irrelevant to the determination of output and employment.

2 A credit crunch is a sharp decline in the supply of credit, especially bank loans.

3 A bank panic or run is characterized by a strong shift in the desired currency/deposit ratio of depositors.

4 An illustration of this comes from explanations of recessions and depressions. In the context of the Great Depression of the 1930s, under the implications of perfect markets, money and shifts in the financial structure would have no impact on real variables, so that the Great Depression would have been solely due to real causes, as specified in the real business cycle theory. Further, models that allow money non-neutrality, but in which all non-monetary assets are identical, explain the monetary contribution to the Great Depression by citing only decreases in the money supply due to the collapse of the banking system as a cause or contributory factor, but otherwise deny credit shortages any role. The financial imperfections hypothesis uses both the decrease in the money supply and the credit restrictions and shortages among the causes of the Great Depression and its duration. There is now significant empirical evidence to support this position.

5 Information imperfections that lead to special financial arrangements for some borrowers, such as in the form of loans rather than bonds, imply that the structure of the financial system determines the sources and uses of funds and can affect real outcomes in the economy.

policy on output and prices/inflation. However, some economists, especially new Keynesians, now believe that because of significant information imperfections a classification of financial assets into three or more assets can explain some monetary policy effects that are not well explained by the two-asset, money and bonds, classification. This chapter examines this issue and specifies a macroeconomic model with money, bonds and credit, which is defined to include loans. Its emphasis on imperfections in financial markets leads to models with a credit channel, in addition to the bond interest rate channel of the transmission of monetary policy on aggregate demand. In some models, information imperfections lead to *credit rationing* (Stiglitz and Weiss, 1981).

For intuition, it would be useful first to review the main characteristics of the non-monetary financial assets in the economy. Bonds in the real-world financial markets are marketable in both primary and secondary markets and their buyers and sellers incur brokerage costs. Marketed bonds do not need any specific collateral, other than the assets of the firm if it were to become bankrupt, nor do they need the credit-worthiness of the issuer of the bond to be established on a one-to-one basis. In terms of the returns on them, their coupon payments and maturity dates are set when they are issued and their price at their maturity date is known in advance. Bonds can be long-term or short-term. The latter, when issued by firms, are called "commercial paper."

A useful definition of credit is short-term debt, including loans of various types, which needs to be rolled over after a short period, so that its amount can be taken to be variable during the short-run. This definition places commercial paper, along with loans and trade credit, among the components of credit. In line with this definition, this chapter defines "bonds" as marketable debt instruments other than commercial paper of firms, and all bonds issued by the government. Of the components of credit, bank loans (including lines of credit) are made directly to customers by financial institutions[6] (designated in this chapter as "banks") and trade credit is extended by firms directly to their customers. More so than bonds, loans and trade credit depend on the direct (as opposed to market) evaluation by the lender of the credit-worthiness of the borrower, which is determined from the individual circumstances of the borrower as well as the state of the markets and the economy.

In general, borrowing through credit can be arranged at short notice and the terms of some of the credit are such that they can be repaid by the borrower or recalled by the lender at any time.[7] Often, for bank loans and trade credit, the credit interest rate is adjustable by the lender even during the duration that the loan is held. The coupon rate on short-term commercial paper gets adjusted at maturity, which occurs only a short period after its issue. As compared with this short-term variability, the coupon on medium- and long-term bonds cannot be adjusted until their maturity, which does not occur for quite some time after their issue. Although the demand for such bonds can fall, this triggers an increase in their yield in the secondary bond markets, but neither the pre-set time pattern of the coupon rate promised by the issuer nor the amount made available to the issuer usually changes until their maturity.

To create a strong analytical distinction between bonds and credit, this chapter assumes that these adjustments in the coupon rate and the amount lent through bonds can occur in the

6 Kashyap and Stein (1994) show that banks in the USA dominate loan financing and the financing of small and medium-sized firms.

7 Given this feature, for loans, the lender can re-assess the credit-worthiness of the borrower at any time and reduce or recall the loan or ask for additional collateral.

long-run but not in the short-run, so that the amount of funds available to firms from bond issues is given for short-run macroeconomic analysis. By comparison, our assumption is that both the amount of credit (including loans and commercial paper) made available to firms and the credit interest rate can be varied in the short-run. In terms of correspondence with the real-world financing arrangements, the preceding strong distinction provides some guidance on where different types of bonds should be slotted. Since our concern is with credit to the private sector, short-term corporate bonds[8] ("commercial paper," including "finance paper") need to be slotted in "credit," along with the loans made directly by banks to firms, since both provide short-term financing to firms and their interest cost to the issuer is frequently adjustable. Hence, short-run variability in the quantity and cost to the private sector is our basis for the distinction between credit and bonds, not marketability versus non-marketability; commercial paper is marketable whereas loans are not, but both are being slotted in credit. Therefore, our analytical concept of credit includes loans and short-term corporate bonds, but not longer-term bonds or government bonds. By implication, our analytical concept of bonds only includes medium- and long-term marketed corporate bonds and all government bonds.

Looking now at stocks, stocks/equities in the real-world financial markets do not have a maturity date and most types of stocks do not promise any coupon payment. Existing stocks can only be traded in the secondary/stock market at a price that continually fluctuates. Since their price depends on the expectation of uncertain future dividends, themselves dependent on the uncertain profitability of the firm issuing them, their yield usually has a higher degree of uncertainty than bonds, which have pre-specified coupon payments and a maturity date, and loans. However, unlike loans, stocks do share with bonds the characteristic that they do not need any collateral up front, nor do they need the credit-worthiness of the issuer to be established on a one-to-one basis.

Given these differences, one could argue that macroeconomic models need to treat money, credit, bonds and stocks as distinct financial assets in the sense that no pair has perfect substitution, so that the overall macroeconomic model needs to have three non-monetary assets and their three rates of return. However, doing so increases the analytical complexity of the macroeconomic analysis beyond the simplicity of the IS–LM or IS–IRT models, which have only one non-monetary financial asset and only one rate of return, so that convincing reasons in terms of the relative impact of shifts in the bonds, loans and stock markets on the economy have to be provided to justify the extension of the macroeconomic model to incorporate three non-monetary financial assets.

This chapter follows other studies that incorporate financial imperfections in limiting the classification of financial assets to just three assets, money, bonds and credit, with stocks still lumped in bonds, and builds a simple form of the macroeconomic model incorporating these three assets. Among several good reviews on the credit channel are Bernanke (1992–93), Kashyap and Stein (1993, 1997), Hubbard (1995), Bernanke *et al*. (1999) and Walsh (2003, Ch. 7).

Note that Walras's law allows one of the markets to be omitted from explicit analysis. Following the usual convention in the IS–LM and IS–IRT analyses, this chapter omits the explicit analysis of the bond market, so that the financial markets explicitly analyzed will be those of money and credit.

8 Some types of such bonds (asset-backed bonds) require backing by other securities, which serve as collateral, which is also often needed for loans.

For aggregate demand in the macroeconomic model, this chapter draws on the model in Bernanke and Blinder (1988) for the supply of loanable funds.[9] For the aggregate supply of commodities, this chapter relies on the *indirect production function*, in which *working capital* becomes an input, along with labor and physical capital in production, with physical capital taken to be fixed in the short-run, as in the usual textbook AS–AD model. Working capital plays the role of an input in a monetary economy by facilitating the purchases of inputs (including labor, raw materials and intermediate goods) and allowing the revenue from sales to accrue to the firm with a lag. This chapter adopts the plausible hypothesis that working capital can vary in the short-run. Since working capital is an input in the indirect production function, a decrease in it reduces the purchases of labor and other inputs, and thereby reduces output.[10]

Working capital

Working capital includes all funds that the firm uses to facilitate its purchases of inputs and production, and sales of its output. The firm may obtain its working capital from its retained earnings, equity issues, bond issues and loans (including the use of overdrafts or lines of credit) – or save on the need for working capital by arranging trade credit. In the long-run, the firm can vary each of the sources of working capital. However, for the short-run analysis, the amount of working capital prearranged by the firm through its retained earnings and issue of bonds (including equities) is taken to be predetermined[11] and not variable. Hence, for the short-run of our model, the only component of working capital that is allowed to vary is that obtained through credit.[12]

Motivating the analysis of the credit market

The study of the money and financial markets would be irrelevant for the determination of output, employment and other real variables if money and non-monetary financial assets were neutral. While such a proposition is implicitly held by virtually all macroeconomic models for the hypothetical, analytical long-run, few macroeconomic models imply it for the short-run, especially in the context of uncertainty, adjustment costs and imperfect competition. In fact, the short-term behavior of the economy is clearly such as to convince the public, economic analysts, central bankers and governments that movements in money and credit, and not merely their unanticipated components, alter output and employment. This can be clearly seen from the very active pursuit of monetary policy by many central banks to stabilize the economy's output around its long-run growth path. It is also obvious from the declines in

9 Bernanke and Blinder provide a commonly used macroeconomic model with money, bonds and loans as financial assets. However, their analysis focuses only on the aggregate demand for output. Kiyotaki and Moore (1997) present another loan-based model intended to explain credit cycles. Their models, along with others, are summarized in Walsh (2003, Ch. 7).

10 The asset-backed commercial paper crisis in USA in 2007 illustrates this role extremely well. Cutbacks in the amount made available in this market meant a severe tightening of credit, with a subsequent cutback in production.

11 Remember, bonds in our model have been defined to exclude issues of commercial paper as a way of supplementing working capital in the short-run.

12 Kashyap *et al*. (1993) and Gertler and Gilchrist (1994) found that commercial paper issuance by large firms expanded during periods of tight credit while those by small firms declined, which represents a *flight to quality* (i.e. to lower risk loans) by the lenders.

output and rises in unemployment caused by currency, credit and exchange crises.[13] Money and credit are, therefore, definitely not neutral in the short-run.[14] This needs to be reflected by the macroeconomic models. The IS–LM and IS–IRT models, at best, do this job poorly and need to be modified to achieve more realistic implications on the impact of a shortage of credit on production. That is, the effects of credit are mainly due to variations in its total, rather than to changes in its composition.

Definitions of credit and loans

If we accept that credit should be treated as distinct from bonds, what should be its definition for macroeconomic analysis? Should it be defined as synonymous with bank loans, as in much of the literature on this topic? We choose to define it as all short-term loans to the private sector, whether marketable or not. As such, its major components are trade credit (provided by suppliers of commodities to buyers), short-term corporate bonds, and loans by banks and other lenders. Therefore, under our definition of credit, loans are just one component of credit.

Our choice of credit rather than merely loans as the distinctive non-monetary financial asset is partly due to our emphasis on credit as the variable component of working capital in the short-run and the impact of variations in it on the production of commodities. Another reason is our belief that information imperfections, discussed in the next section, which are usually behind the treatment of loans as a distinctive financial asset, really affect all financial assets, including trade capital, short-term bonds and long-term bonds, rather than merely loans. Admittedly, banks, when making loans on a personal basis, cope with market imperfections in a different way to financial markets, but our belief is that this difference is less important than that between short-run variability (for all forms of credit) versus non-variability (for bonds), which is what is needed to distinguish between short-run and long-run effects. Further, trade credit and bank loans suffer in a very similar fashion from market imperfections. In addition, an advantage of including bank loans, trade credit and commercial paper within the single category of credit is that if changes in one of them offset to some extent changes in another, only the net change in the total becomes relevant for the analysis.[15]

In any case, if one chooses, credit can be replaced by loans in this chapter, without changing the gist of its arguments.

Links between credit and economic activity and between credit and monetary policy: lessons of the 2007 subprime asset crisis for macroeconomic analysis

The 2007 crisis in the subprime asset-backed credit market (ABCM) in the USA, and its impact on housing construction and real economic activity generally in the USA and the

13 To illustrate, in 2007, during the crisis in the subprime financial markets, speeches and reports on its impact on economic activity often carried headings such as "Heading for the rocks: will financial turmoil sink the world economy?"

14 Their non-neutrality was starkly illustrated during the subprime mortgage crisis in the USA during 2007, when the reappraisal of risk on mortgage-backed securities led to the drying up of the demand for short-term securities by financial institutions and caused fears that the shortage of working capital would lead to cutbacks in production and a recession in the US economy, and in the world economy.

15 Several studies report that a tightening of bank loans leads to an increase in commercial paper issued by larger firms.

world, illustrates how shifts in the availability of credit affect real economic activity and how the availability of money affects that of credit, so that neither credit nor money is neutral for real economic activity. House prices in the USA had shot up during 2001 to 2006 to such an extent that the rise was generally considered to be a bubble. This rise had been prompted by shifts in investment after the collapse of stock prices due to the crash in Internet stocks in 2001–02, and by exceptionally low interest rates during 2001 to 2006. As house prices rose, the demand for housing was further boosted by the practice of mortgage lenders to ease the terms for obtaining mortgages: some mortgages were for 100 percent of the purchase price of a house, even to customers who did not have the income to cover the projected monthly mortgage payments. This practice did not pose a serious problem as long as house prices continued to rise sufficiently fast, since the price increase provided a cushion for both buyers and lenders. However, house prices began to stabilize and then to fall in 2006. Further, interest rates began to rise. The result was an increase in mortgage defaults and a heightened perception of the riskiness of such mortgages.

Collections of the initial mortgages and other credit, such as installment credit on purchases of cars, were bundled into marketable securities, with the former serving as collateral for the latter. The securities usually had terms of 30 to 60 days, similar to that on Treasury bills, and were called asset-backed commercial paper. Since they carried a higher yield than Treasury bills but seemed to be equally liquid and safe investments, they proved attractive to many lenders as a component of their portfolios. They were bought by (production) firms, as well as by commercial and investment banks, etc. Once house prices in the USA began to fall and the risks of mortgage default rose, the riskiness and eventually illiquidity of such asset-backed commercial paper became apparent, so that the demand for them fell drastically, while the interest rates on them rose. With many issuers of such bonds unable to roll them over, the risk of default in this market rose significantly. This reduced the flow of funds for working capital to firms, affecting their production. The possibility of such a default also affected the holders of such securities, reducing their liquidity and profitability. Since many banks and other firms in the USA and many other countries held such securities, the crisis in the subprime (high-risk) market threatened to become a general financial crisis, which in turn threatened to deplete the amount of working capital for firms in production and send the USA and the world economies into a recession.

The subprime crisis could be viewed as one of the short-run liquidity, rather than long-term solvency, of most financial institutions holding the mortgage-backed securities. As part of attempts to limit the impact of the subprime crisis, central banks in the USA, Europe and other affected countries pursued an aggressive expansionary monetary policy by increasing the money supply through open market operations, cutting discount rates and openly calling for commercial banks to borrow funds as needed from the central bank. In the USA, the Fed increased the money supply very significantly, cut its discount rate by 1 percent (and openly encouraged bank borrowing at this rate), following this by a cut in the federal funds rate by ½ percent, with subsequent cuts occurring gradually. In spite of the aggressive actions taken by several central banks, the fallout of the subprime crisis for real economic activity remained uncertain for quite some time: while the projections of the real growth rates of the world economy were cut, such a reduction was nominal for the optimistic scenarios and drastic for the pessimistic ones, with the latter forecasting a severe world recession. However, it was soon clear that, by reducing the working capital of firms, the subprime crisis had the potential to cause a recession in output and that aggressive expansionary monetary policy did moderate but not eliminate this effect.

The day-by-day and week-by-week pronouncements of central banks and economic analysts during this period did also highlight the very considerable uncertainty of the extent of the need for monetary policy and the difficulty of formulating the appropriate policy in a timely fashion.

Briefly, firms habitually rely on credit to conduct their production, so that a credit crisis, by reducing the credit made available to firms, is likely to reduce their production in the short-run. If the credit markets do not provide the needed liquidity in sufficient amounts and in a timely manner, a credit crisis tends to evolve into an economic one, in which the economy goes into a recession. The standard IS–LM, AD–AS model does not provide a link from credit to production, so the subprime crisis cannot be explained by using it. This chapter tries to provide a model that links credit to the production activities of firms and also links credit to the money supply.

16.1 Distinctiveness of credit from bonds

16.1.1 Information imperfections in financial markets

The information that borrowers and lenders bring to their exchanges and the ability of borrowers to change the probability of default play an important role in the nature of credit contracts, the ability of credit markets to match borrowers and lenders, and the role played by the interest rate and credit rationing in the allocation of credit among borrowers. Given that there is a risk of default by a borrower, risk-averse lenders base their decision to lend on the expected return while borrowers need only base their decision to borrow on the interest rate that they have to pay. While the terms of the loan usually guarantee the latter, they do not guarantee the probability of not defaulting. Credit-rationing models imply that lenders will not raise interest rates beyond the point at which their expected return begins to decline, even though some (higher risk) borrowers are willing to pay higher interest rates (Jaffee and Russell, 1976; Stiglitz and Weiss, 1981). In this case, the credit market "equilibrium" would be characterized by an excess demand for credit, so that there would be quantity rationing (through the amount lent), in addition to price rationing through the interest rate.

"Credit rationing" occurs if, at the going interest rate, lenders supply a smaller amount than the borrowers want to obtain. In addition, among borrowers willing to pay the going interest rate, some borrowers receive credit while others do not. In fact, the latter may be willing to pay a higher interest rate but may still not get the credit. Hence, there is an unsatisfied demand for credit that is not eliminated by a rise in the interest rate. Consequently, equilibrium in the credit market is characterized by the equilibrium interest rate and credit rationing (Jaffee and Stiglitz, 1990; Stiglitz and Weiss, 1981).

Information imperfections[16] in financial markets occur because of adverse selection, moral hazard, and monitoring and agency costs – and because of the undeveloped and segmented nature of the financial sector of some economies. Under uncertainty, different borrowers can have different probabilities of repayment (versus default). If the lender can observe these

16 Akerlof (1970) illustrated the role of information imperfections in the used car market by assuming that used car sellers know the quality of the car but buyers do not. The latter view lower prices set by sellers as indicative of poorer quality, so that lowering the price does not increase demand. Therefore, in the used car market, an excess supply of cars need not be cleared by a lower price. In fact, there may be no price at which supply equals demand.

probabilities and if they are not affected by the decision to lend, lenders can accurately rank the borrowers on the basis of their expected return and make loans to those who will yield the highest expected return. In practice, lenders cannot actually observe the borrower's probability of repayment or rely upon this probability remaining invariant. To illustrate, start with equilibrium in the credit market at the current *expected* return and assume that there is an unsatisfied demand by borrowers at this equilibrium. Given this unsatisfied demand at the equilibrium interest rate, suppose that the lenders were to increase the interest rate. At the higher rate, some borrowers, including those with less risky projects, might no longer find it profitable to borrow, while those with riskier projects may still be willing to borrow. This shift among the borrowers in the probability of repayment as the interest rate rises will, from the lender's viewpoint, represent *adverse selection*.[17] This shift in the mix of borrowers to those with lower probabilities of repayment will lower the lender's expected return, even though the credit interest rate has risen, so that, beyond a critical credit interest rate,[18] lenders will not find it to their advantage to raise the interest rate, or increase their loans, in spite of any unsatisfied loan demand. Rather, they will ration their funds among the borrowers on some basis, such as adequate and acceptable liquid collateral or their balance sheet position, other than merely the willingness to pay the going interest rate. This credit interest rate, then, is the equilibrium rate, though with rationing. Hence, borrower heterogeneity and adverse selection results in credit rationing at the equilibrium credit interest rate because lenders find it unprofitable to raise the credit interest rate, even in the face of the excess demand for loans.

Moral hazard arises when borrowers can choose among projects with different degrees of risk and lenders cannot monitor this choice. Higher credit rates of interest can entice borrowers to use the funds for riskier projects, which will reduce the lender's expected return. Just as in the case of adverse selection, moral hazard results in credit rationing at the equilibrium interest rate.

Further, under the imperfection of information available to the lender, once a loan has been arranged, the borrower may have an incentive to under-report the success of the project – and some borrowers may have an incentive to default on the loan. To offset this tendency, as well to contain the rise in risks due to adverse selection and moral hazard, lenders can resort to monitoring, on their own or through an agent, the projects financed with the loans. However, doing so involves some costs, usually labeled as *monitoring and agency costs*. These costs reduce the return, as well as the expected return, from the loan to the lender, while the cost to the borrower of the loan remains at the interest rate. Such costs would be especially applicable where the payment to the lender is positively related to the success of the project financed with the borrowed funds, so that the borrower would have an incentive to under-report the profitability of the project.[19] Agency and monitoring costs do not arise where the firm uses its internal funds but do arise when it borrows, so that they raise ("drive a wedge") the relative costs of external versus internal funds, with

17 In the used car example in Akerlof (1970), as used car prices fall, sellers of poorer quality cars ("lemons") will be more willing to sell their cars than the sellers of better quality ones, so that there will be adverse selection by the sellers in the quality of the cars offered for sale.

18 This critical rate is one at which the fall in the probability of repayment is more than proportional to the rise in the loan rate. Therefore, the lenders' expected return function has a local maximum at the critical loan rate.

19 Among the contributions to the effect of adverse selection, moral hazard, monitoring and agency costs on credit markets are Jaffee and Russell, 1976; Jaffee and Stiglitz, 1990; Stiglitz and Weiss, 1981; Williamson, 1987 and Bernanke and Gertler, 1989).

this differential in costs increasing with the proportion of external to internal funds. Since recessions worsen firms' balance sheets and also reduce the availability of internal funds, they increase agency costs, thereby reducing investment, thereby worsening the recession. Hence, imperfect information in credit markets can amplify the impact of shocks to the economy.

From the borrowers' practical perspective, the impact of information imperfections and transactions costs means that either some firms are unable to raise funds in the bond market or they find it less costly to borrow from banks (henceforth, this term includes other providers of credit) rather than through the bond market, so they have to rely on loans and trade credit for all or some of their borrowing needs. In practice, while all firms rely to some extent on credit for short-term financial needs, small and medium-sized firms – as well as households who need to borrow – needing relatively small amounts do so to a much greater extent. They get barred from the bond markets because of the high transactions costs of bond issues. Firms can also get barred from the bond markets because of high monitoring costs, which require direct assessment by a lender before it will lend. Such firms have to rely on loans and trade credit for their external finance. On the lender's side of the credit market, providers of loans and trade credit have an advantage over the bond market because of their better ability to individually assess and monitor the creditworthiness of the borrower.[20] They also have a cost advantage in being able to lend relatively small amounts. In any case, they can effectively offer lower costs of credit than the bond market to some types of borrowers, especially small and riskier ones. This is also so for borrowing by consumers for purchases of durable goods.

To conclude, while information imperfections can play an important role in both credit and bonds markets, their role is stronger in credit markets since credit is short-term and needs to be renewed often. They strengthen the distinction between bonds and credit beyond merely the difference in marketability and provide a justification for the separation between bonds and credit markets for macroeconomic analysis. This distinction should be positively related to the degree of information imperfections. It would be more intense for small than for large firms. It would be also more intense, on the whole, in financially underdeveloped than in developed economies, so that production units in the former would have to rely more on internal sources of funds (often savings of owners and close relatives) and, when they do, rely more on small loans and trade credit provided under informal arrangements than on (marketed) bond issues.

Impact of monetary policy on firms' balance sheets and borrowers' creditworthiness

Monetary policy has both direct and indirect effects on firms' financial positions and therefore on their ability to obtain external funds.[21] Since firms usually have some variable-rate credit, a tight monetary policy, by raising interest rates, increases their interest payments. In addition, the rise in the interest rates may also cause a decrease in the prices of their assets and reduce the value of their collateral. These direct effects of monetary policy on the

20 Over time, the bank lending to a firm acquires an information advantage relative to other banks, with other lenders left with an information disadvantage, so that the firm becomes dependent on a particular bank.

21 Empirical evidence showing that internal finance is cheaper than external finance and that balance sheets matter is now quite well established (Bernanke, 1992–93).

firms' balance sheets are supplemented by an indirect effect, which occurs because a tight monetary policy reduces the demand for their products and their revenues. Therefore, both the direct and indirect effects of the tight monetary policy tend to reduce the creditworthiness of borrowers.

The preceding discussion provides two subdivisions of the way the credit markets affect the economy. These are:

- *The pure credit channel*. This affects the amount supplied of credit, for given creditworthiness of borrowers, as well as its interest cost. A component of the pure credit channel is the *bank-lending channel*, which emphasizes the special nature of bank credit and the role of banks in the economy.
- *The balance sheet channel*. This affects the demand for credit and the creditworthiness of borrowers.[22]

Market imperfections and the bank lending channel

Monetary policy affects the supply of bank loans and the banks' demand for short-term corporate paper. Banks rely mostly on deposits for virtually all of their funds. A contractionary monetary policy reduces these deposits. While banks are nowadays able to borrow directly in financial markets by issuing marketable liabilities, such as certificates of deposit and short-term bonds, these instruments are not perfect substitutes for deposits since they bear a higher interest rate than bank deposits, nor is the demand for such bank liabilities perfectly elastic.

Looking at the portfolios of "commercial banks" (using this term for a wide variety of financial institutions), the two distinct types of assets, in addition to reserves, held by banks are bonds and credit (which under our definition includes loans and commercial paper). In general, commercial banks usually do not hold stocks. A considerable proportion of the loans made by banks is often to small and medium-sized firms or consumers, who choose to take loans from banks which specialize in monitoring and enforcing contracts, because it is less expensive for them to borrow on a one-to-one basis than in the bond and stock markets. Large firms also finance a part of their working capital by short-term finance paper, which carries lower interest costs than longer-term bonds.

Therefore, the core aspects of the bank-lending channel are the lack of close substitutes for deposits on the liability side of the banks' balance sheets and the lack of close substitutes for credit (i.e. bank loans and short-term paper) for borrowers. As a result of the former, contractionary open-market operations shrink the banks' deposit base. If they try to offset this shrinkage by borrowing in the bond and equity markets, this increases the relative cost of their funds. Hence, banks' response to the shrinkage of their deposit base reduces the supply of loanable funds and raises the rates charged. On the borrowers' side, from the balance sheet mechanism explained above, the worsening of their balance sheets means that some of the borrowers will be squeezed out of the credit market while others will have to pay higher rates.

The relative amounts and significance of credit versus bonds is a relevant consideration in deciding whether a separate credit market needs to be considered explicitly for

22 Since recessions reduce sales and worsen balance sheets, the worsening of the balance sheet will reduce the affected firm's access to credit and raise its cost, even though bond interest rates may have fallen.

macroeconomic analysis. Empirically, borrowing in the form of loans rather than bond issues is very much more important in countries with underdeveloped bond and stock markets than in the financially developed economies. Even for the latter, some economists claim that loans are significant enough in size and macroeconomic impact to be explicitly modeled apart from the bond market (Kashyap and Stein, 1993, 2000), although the usual mode of macroeconomic analysis chooses not to do so for such economies and uses the IS–LM or IS–IRT model.

Formally, the distinction between bonds and credit is not needed if either the borrowers or the lenders are indifferent between them: from a macroeconomic perspective, the separate modeling of bonds and credit is needed only if loans and other forms of credit are subject to quantity constraints[23] (i.e. rationed) on some basis other than their cost, and/or the credit rates of return/interest are different from and not perfectly correlated with the bond interest rate. The greater the slippage between the two rates or the more important the rationing constraint, the more relevant becomes the need for a separate analysis of the credit market. Loan rationing by banks and the differential in credit and bond interest rates are likely to be greater if the financial markets are fragmented, as often occurs in financially underdeveloped economies, or regulated. Examples of the latter are legal limitations on interest rates[24] and controls on amounts or proportions of bank portfolios allocated to credit to specific sectors of the economy. Many economists believe that in economies with competitive and efficient financial systems, the distinctive and significant role of the credit channel *vis-à-vis* the bond channel does not arise from the credit supply side, and therefore not from the role of banks in lending, but from the demand side, so that the borrowers' creditworthiness plays an important role. That is, the supply side of the credit channel, especially the lending one, is not significant while the balance sheet one is.[25]

Given our intent of formulating the simplest model incorporating credit as a distinctive financial asset, we stylize the preceding remarks in the following manner. Since different firms rely on credit to somewhat different degrees for their working capital, we attribute to the representative firm the average amount of funds raised through credit by all firms. This firm would, then, have a part of its funds raised through credit, with the remainder being from retained earnings or raised through bonds (including stocks), but the short-run variations in its working capital will come only from credit. Even for credit, for reasons of

23 There is effective quantity rationing in the loan market if banks lend different amounts at any given loan rate. Such rationing does not occur in perfectly competitive markets, but the loan market, with one-to-one lending and customary monopolistic relationship established over time between a bank and its borrowers, is likely to display effective quantity rationing in the short-run.

24 A historical illustration of this from the USA is Regulation Q, which limited the interest rate that banks could pay on their deposits. Such limits were eliminated in the 1970s. However, they remain common in many less developed economies.

25 Ashcroft (2006) uses affiliation with a bank holding company as a proxy for financial constraints to study the behavior of banks in the USA. While small banks tend to react strongly to monetary policy changes, their behavior is counterbalanced by those of other banks, so that on average the distinct effect of monetary policy through bank loans is insignificant. Ashcroft and Campello (2007) examine loans extended to small local businesses by small subsidiary banks with the same holding company but operating in different geographical areas of the USA. Their findings show that cross-sectional variations are not due to the response of bank lending to monetary policy but depend on the creditworthiness of borrowers. These studies therefore support the balance sheet channel as against the bank lending channel for impact of monetary policy.

analytical simplification, our analysis will ignore short-run variations in trade credit. Our stylized definitions and assumptions for short-run analysis are as follows:

- "Banks" are defined as being retail banks and include moneylenders as well as other financial institutions that provide loans, usually on a one-to-one basis. For short-run analysis, banks are assumed not to hold bonds and stocks in their portfolios or, alternatively, these holdings are treated as constant in the short-run but variable in the long-run.
- Credit constitutes a positive fraction of the working capital of firms. A decrease in credit decreases the amount of working capital.
- While bonds and retained earnings of firms provide some fraction of the working capital of firms, the amount raised through them for working capital cannot be changed in the short-run, though it can be varied in the long-run.
- The working capital of firms is an argument in the indirect production function of the representative firm, with output a positive function of working capital.

16.2 Supply of commodities and the demand for credit

The representative firm engages in production and has to pay for its inputs, including labor services,[26] prior to sales. The funds used for such payments are the firm's "working capital."[27] If the firm holds inadequate funds relative to the payments to be made, it has to reduce its purchases of inputs, which reduces output. While such usage can be introduced through a cash-in-advance constraint, we choose the different route of introducing working capital in the indirect production function (see Appendix B to this chapter).[28] As Appendix B shows, working capital, in addition to the firm's use of labor and physical capital, is a determinant of output, so that the indirect production function of the representative firm can be specified as:

$$y = y(n, \overline{K}, k^w) \tag{1}$$

This production function is assumed to be twice differentiable, with the first derivatives being positive and the second negative. n is the number of workers, k^{w} is real working capital and $\overline{K}$ is the exogenously given (for short-run analysis) physical capital stock. Henceforth omitting $\overline{K}$ from the production function, this function becomes:

$$y = y(n, k^w) \tag{2}$$

Our assumptions above were that part of the working capital is raised through some combination of bonds, retained earnings and loans. Since the use of working capital arises from the need to pay labor (and for other inputs), profit maximization by the firm implies that the demand for working capital will depend on the real wage rate and on the real interest

26 The following analysis uses the number of workers employed as the proxy for all inputs, and the wage rate as the proxy for their cost.

27 The notion of working capital is not new in the literature. To illustrate, Keynes (1937) used loans as a constraint on investment undertaken by firms. It has also been used as a variable in the production function in some studies.

28 This is justified by the argument that if the working capital is inadequate, the firm has to take some of its workers from production and divert them to facilitating purchases of inputs and sale of output.

rate levied on the funds raised for this purpose.[29] Hence, the demand for working capital is given by:

$$k^{w,d} = k^{w,d}(\underset{-}{w}, \underset{-}{r^L}) \tag{3}$$

where r^L is the interest rate on credit, which is the short-run variable part of working capital. The signs under the variables are those of the respective partial derivatives. $\partial k^{w,d}/\partial w$ is negative since an increase in the wage rate reduces the employment of labor, which reduces the need for working capital. Since r^L is the interest cost of the working capital financed through credit, $\partial k^{w,d}/\partial r^L$ is negative.

One of our stylized assumptions above was that, in the short-run, changes in the supply of or demand for bonds do not affect the amount of funds made available to a firm, nor do such changes affect the interest cost of the pre-existing bonds issued by the firm, so that, for short-run analysis, this cost becomes part of the fixed cost of the firm. Therefore, short-run production analysis only needs to take account of the real cost of credit but not of the bond rate, thereby making output, the demand for labor, real working capital and credit functions of the real wage rate and the real credit interest rate.

Supply function for commodities

For the short-run indirect production function (2), the Euler conditions for profit maximization by the firm yield the demand for labor as $n^d = n^d(w, r^L)$. Assuming the simple labor supply function as $n^s = n^s(w)$, labor market equilibrium implies that $n = n(r^L)$. Substituting this function in the production function yields the supply of output as:

$$y^s = y^s(r^L) \tag{4}$$

The reason why $\partial y^s/\partial r^L \leq 0$ is that an increase in r^L reduces the working capital used by the firm, which, given the specification of the indirect production function, reduces its output. $\partial y^s/\partial r^L = 0$ if the working capital held by the firm exceeds the maximum needed for purchases of labor and other inputs. The analysis of this maximum is illustrated in Appendix A.

Demand function for credit

Further, for short-run analysis, profit maximization by the firm with the posited short-run production function implies, as shown in Appendix B, that the bond rate can also be omitted from the demand function for loans, so that:

$$L^d/P = \psi(\underset{-}{w}, \underset{-}{R^L}) \tag{5}$$

where L^d is the nominal demand for credit and $\psi(w, R^L)$ is the real demand. Since we do not wish to explicitly model the labor market in this chapter, we replace the wage rate by output. With an inverse relationship between w and y, rewrite (5) as:

$$L^d/P = \psi(\underset{+}{y}, \underset{-}{R^L}) \tag{6}$$

29 Appendix B expands on the nature and role of working capital in production, and its cost.

Note that, as shown in Appendix A, the firm's demand for credit has an upper limit defined by its overall need to finance the purchases of inputs. However, its actual demand for credit will be less than this limit since some of the working capital needs will have been prearranged by other sources, such as bonds and retained earnings.

16.3 Aggregate demand analysis incorporating credit as a distinctive asset[30]

16.3.1 Commodity market analysis

The commodity market of our open economy model is specified in the customary manner and encompassed in the standard IS equation and curve. Drawing on the familiar analysis of the IS market for the open economy, the general form of the IS equation for the open economy is:

$$y = y(\underset{-}{r}, \underset{-}{r^{\mathrm{L}}}, \underset{-}{P}) \tag{7}$$

where P is the domestic price level. In this function, the dependence of y on both r and r^{L} occurs because the real cost of raising funds for investment and working capital is given by r for the proportion raised by the representative firm through bonds (with the same rate acting as the opportunity cost of using retained earnings) but by r^{L} for the proportion raised through credit.[31] The dependence of y on P in the open-economy IS equation occurs because of substitution between domestic and foreign commodities (see Chapter 13).[32] The IS curve related to (7) has the usual negative slope whether r or r^{L} is on the vertical axis, and shifts with a change in P.

16.3.2 Money market analysis

Money market equilibrium

Assume that households and (production) firms (as against banks and moneylenders) do not have the choice of making loans[33] but can hold, among their assets, bonds (which include savings deposits, money market mutual funds, stocks, etc.) as the alternative to holding money. Assuming, as usual, an exogenous pre-determined level of financial wealth, FW_0,

30 The basic structure of the commodity and money markets, and of the specification of bank loan supply, draws heavily on Bernanke and Blinder (1988). These authors point out that the standard IS–LM model incorporates an asymmetrical treatment of the banks' liabilities and assets. The former, through checkable deposits, are incorporated in the IS–LM model while the asset side is ignored. Their model modifies the IS–LM component of macroeconomic models, but not the output supply analysis of these models.

31 If the credit market has quantity rationing in addition to that exercised by the loan rate, so that different amounts of credit are made at a given interest rate and the amount lent affects investment or consumption, the amount loaned becomes an additional argument in the IS function.

32 This occurs when purchasing power parity (PPP) is not imposed on the model. PPP rarely holds in the economy even over long periods, and the reversion to it is remarkably slow. Its assumption is inappropriate in a short-run model.

33 This assumption could be questionable for economies with large informal financial markets.

to be allocated to money and bonds, the standard analysis of the demand for real balances (m^d) specifies the money demand function (see Chapter 13) as:

$$m^d = m^d(\underset{+}{y}, \underset{-}{R}; \underset{+}{FW_0}) \tag{8}$$

where m^d is real money demand, y is real national income and R is the nominal bond rate. For perfect capital markets, R and r are related by the Fisher equation:

$$(1+R) = (1+r)(1+\pi^e) \tag{9}$$

where π^e is the expected rate of inflation. Assuming π^e to be exogenously set (for simplification, at zero) or choosing to ignore it because our analysis will be a comparative static one, $(1+\pi^e)$ is suppressed in our further analysis, so that we treat R and r as being equal (Bernanke and Blinder, 1988). Hence, the preceding money demand equation becomes:

$$M^s = P \cdot m^d(y, r; FW_0) \tag{10}$$

For a given money supply M^s, money market equilibrium specifies the usual LM equation as:

$$M^s = P \cdot m^d(y, R; FW_0) \tag{11}$$

Defining M^s narrowly as the sum of currency C in the hands of the public and bank deposits D, $M^s \equiv C + D$. However, we also have M0 $\equiv C + (RR + FR)$, where M0 is the monetary base M0. Therefore, the money supply is determined (see Chapter 10) as:

$$M^s \equiv \frac{\text{M0}}{\left[\dfrac{C}{M} + \dfrac{(RR+FR)}{D} - \dfrac{C}{M} \cdot \dfrac{(RR+FR)}{D}\right]} \tag{12}$$

where C/M is the currency ratio, RR/D is the required reserve ratio, FR/D is the free reserve ratio and $(RR+FR)/D$ is the (actual) reserve ratio. Hence, the form of the LM equation that is more appropriate for the pursuit of monetary policy than one with an exogenous money supply is:

$$\frac{\text{M0}}{\left[\dfrac{C}{M} + \dfrac{(RR+FR)}{D} - \dfrac{C}{M} \cdot \dfrac{(RR+FR)}{D}\right]} = P \cdot m^d(y, r) \tag{13}$$

The central bank controls the monetary base M0 and the required reserve ratio RR/D, while the public determines the currency ratio C/M and the banks determine the free reserve ratio FR/D.

Supply of bank loans

In the short-run, the supply of credit (loans, short-term commercial paper, trade credit, etc.) comes from various sources: it is provided by financial institutions, markets in short-term corporate bonds and suppliers to buyers of commodities. Of these, it is easiest to model

the supply of credit by banks, so that the supply of credit is usually modeled as the supply of loanable funds by banks. Except for small other items, banks hold required reserves (RR), free reserves (FR), bonds (B) and credit (including loans) L on the asset side of their balance sheet, and have deposits D as liabilities, so that their balance sheet provides the identity:

$$B^{\mathrm{d}} + L^{\mathrm{s}} + FR \equiv (1-\gamma)D$$

where γ is the required reserve ratio RR/D and the superscripts d and s refer to demand and supply respectively. Under our definitions above, B refers to longer-term bonds and L (for credit) includes loans and short-term commercial paper. For a given monetary base M0, an increase in the banks' desired or required reserves or in the currency/demand deposit ratio reduces the amount banks allocate to credit and bonds and raise the loan and bond rates. An increase in the monetary base increases the money supply, bank reserves, the demand for bonds and the supply of credit. The latter reduces the return on both these assets. The allocation of funds between bonds and credit depends on their relative risks, which, as discussed earlier, depend on the creditworthiness of the borrowers. An increase in the relative riskiness of credit, as caused by a recession or other adverse shocks to the balance sheets of the borrowers, will decrease the amount of credit provided by banks and increase that of bonds, representing a "flight to quality."

In view of the banks' balance sheet identity above, the impact of monetary policy on their assets depends on its impact on their budget constraint, which depends on their ability to substitute between the various types of their liabilities. Among these liabilities are demand and savings deposits, which are initiated by the public, and other short-term liabilities (such as marketed certificates of deposit (CD)) initiated by the banks. If banks can protect the sum of their liabilities from restrictive monetary policies, for example by offsetting a policy-induced fall in the monetary base and deposits by increasing the sales of their marketable liabilities domestically or by borrowing abroad, they will eliminate the impact of the monetary policy on the credit provided by banks – and, through the bank-lending channel, on the credit interest rate. This is more likely to occur in financially developed economies than in underdeveloped ones. Our assumption on this is that the banks cannot do so even in the financially developed economies, so that restrictive monetary policies reduce banks' deposits and their supply of credit, and raise the credit interest rate.

16.3.3 Credit market analysis

Supply of credit

Portfolio selection behavior by deposit-taking banks implies that their holdings of free reserves, credit and bonds depend on the nominal rate of return R on bonds, the nominal rate of return R^{L} on credit and on their "scale variable" $(1-\gamma)D$, so that the banks' supply function for credit is:

$$\underset{}{L^{\mathrm{s,B}}} = \underset{+}{\lambda}(\underset{+}{R^{\mathrm{L}}}, \underset{-}{R})(\underset{+}{1-\gamma})(\underset{+}{D}) \tag{14}$$

where $L^{\mathrm{s,B}}$ is the supply of funds by commercial banks in credit markets.

The financial system in virtually all economies has a complex, layered structure, with some financial institutions borrowing from others, and the latter borrowing from still others.

If we place commercial banks, which take deposits from the public, at the apex[34] of a triangle designating this system, we can envisage lower layers of financial institutions borrowing from those above it. The earlier analysis of market imperfections, leading to adverse selection, moral hazard, and monitoring and agency costs, applies to such borrowing by the lower layers from higher layers of financial institutions, so that the allocation of such credit among the layers is characterized by both the credit interest rate and credit rationing. It will also depend on the net worth of the borrowing units and the perceived riskiness of lending to them. These will in turn depend on the composition of their portfolios of assets. At the bottom of the triangle depicting the layers of financial institutions, the assets of such institutions will consist of credit provided to production firms and households, so that the riskiness of credit provided to financial institutions will depend on the riskiness of credit to firms and households. Hence, the supply of credit by the financial system cannot be analyzed purely from the perspective of the balance sheet of the commercial banks, so that even if the monetary base does not change, the supply of credit could vary with shifts in the riskiness of credit along the layers of the financial system.

Somewhat distinct from the layers of financial institutions is the layered structure of credit itself. In developed financial systems, credit is more than loans extended by a commercial institution that has a fair knowledge of the risks in them and holds them to maturity. The initial credit or loans can be bundled in various ways and resold as marketable securities, so that the underlying risk can become obscured, even to other financial institutions. Further, the marketable securities thus created can, in turn, be used as collateral/backing for borrowing by their holders, thus leading to a multiple creation of credit that is not captured by the bank deposit creation process, as well as to further lack of transparency of risk through the credit layers. As discussed above in this chapter, the asset-backed commercial paper crisis, originating in the USA in 1970 and spreading to many other developed countries, highlighted such multiple creation of credit, along with the potential for obscured risks along the layered structure of the credit system.

Therefore, the supply of credit is more than simply a matter of loans being extended to ultimate borrowers by the banks themselves, with the risks in such loans evaluated reasonably well by the lending bank. With risk obscured in the layers of the credit supply, shifts in the creditworthiness of borrowers can reverberate through the credit supply chain and impact on both the credit interest rate and, sometimes more so, on the quantity of credit extended at the various layers,being eventually reflected in changes in credit rationing to production firms and households. Therefore, we specify the supply of credit as:

$$L^{s} = \underset{+}{L^{s}}(\underset{+}{\lambda}(\underset{+}{R^{L}}, \underset{-}{R})\underset{+}{(1-\gamma)D}, \rho) \tag{15}$$

where L^{s} is the overall supply of funds in credit markets and ρ is a (very simplistic) proxy for risk, credit rationing and structure of the credit system, so that shocks to ρ can come from shocks to any of these factors. If ρ does not change, fluctuations in credit supply will mainly reflect fluctuations in the supply of bank loans.

34 In a system where the central bank is the lender of last resort, the central bank is really at the apex.

Demand for credit

The demand function for the nominal value L^{d} of credit was derived earlier from production analysis as:

$$L^{\mathrm{d}} = P \cdot \underset{-}{\psi}(\underset{-}{R^{\mathrm{L}}}, \underset{+}{y}) \tag{16}$$

As discussed earlier, under imperfect information, the creditworthiness of the borrower is a significant element in addition to the promise to pay the interest payments on loans and determines the expected return to the lender, as encompassed in the balance sheet channel discussed earlier. For simplicity, creditworthiness may be taken to be proxied by current output y, though it will also depend on the expected future output, so that creditworthiness is not entered as an additional argument in the credit demand function.

Credit market equilibrium

From the above, the credit market equilibrium condition is:

$$P.\psi(R, y) = L^{\mathrm{s}}[\lambda(R^{\mathrm{L}}, R)(1-\gamma)D, \rho] \tag{17}$$

We will designate this equilibrium equation for the credit market as the "*CC equation*."[35] Since it was assumed earlier that the expected inflation rate is exogenously given and, for convenience, set at zero, $R^{\mathrm{L}} = r^{\mathrm{L}}$ and $R = r$. Hence, the CC equation can be rewritten as:

$$P.\psi(r^{\mathrm{L}}, y) = L^{\mathrm{s}}[\lambda(r^{\mathrm{L}}, r)(1-\gamma)D, \rho] \tag{17'}$$

Note that this equilibrium condition hides many salient aspects of the credit market in ρ and only explicitly captures the bank-lending channel under the *ceteris paribus* clause of no alteration in the riskiness and quantity of credit.

16.3.4 Determination of aggregate demand

For comparative static analyses, the IS equation (7), the LM equation (13) and the loan market CC equation (17) specify the three equations needed for deriving aggregate demand.[36] They can be solved for ρ, r (or for R, given π^{e}) and aggregate demand y^{d} in terms of the values of the exogenous monetary policy variables γ and M^{s}, and the fiscal policy variables.

35 Bernanke and Blinder (B&B) (1988) do not derive the demand for loans but instead assume a priori the loan demand function to be:

$$L^{\mathrm{d}} = \underset{-}{L^{\mathrm{d}}}(\underset{-}{R}, \underset{+}{R^{\mathrm{L}}}, \underset{+}{y})$$

Hence, the B&B loan market equilibrium condition is:

$$L^{\mathrm{d}}(r, r^{\mathrm{L}}, y) = \lambda(R^{\mathrm{L}}, R)(1-\gamma)D$$

36 This model of aggregate demand would be very similar to that of Bernanke and Blinder (B&B) (1988) if credit were defined as bank loans and our emphasis on quantity rationing and the layered structure of credit and financial institutions, designated by ρ, were ignored.

For this purpose, first combine the IS and credit market equilibrium equations by eliminating R^{L} to derive the "IS–CC equation." The general form of this equation is:

$$y^{\mathrm{d}} = y^{\mathrm{d}}(r, P, (1-\gamma)D; g, \rho) \tag{18}$$

where g stands for the fiscal policy variables. In (18), $\partial r/\partial y < 0$ since a rise in r reduces investment, as for the IS curve, so that the IS–CC curve has a negative slope in the (r, y) space. However, note that while the IS curve is independent of the money market, the IS–CC curve shifts with changes in the monetary base M0, the currency ratio C/D and the required reserve ratio γ, which together determine deposits D. Shifts in this curve can also occur because of shifts in M0 and ρ.

If credit and bonds are perfect substitutes[37] for either the borrowers or the lenders, so that $R^{\mathrm{L}} = R$ and the supply of credit is not rationed on a basis other than their rate of return, or if commodity demand does not depend on the loan interest rate, the IS–CC equation degenerates to the IS one, so that its combination with the LM equation makes the model the standard IS–LM curve for determining aggregate demand. However, intuitive and empirical information on the economy seems to indicate that these conditions are not likely to be met for the financially developed economies,[38] and even less so for the financially underdeveloped economies. Further, our model assigns distinctive short-run roles to credit and bonds in the production sector, so that they cannot be perfect substitutes in our model.

16.4 Determination of output

In the Blinder and Bernanke (B&B) (1988) analysis, variations in aggregate demand cannot change output unless there are imperfect price adjustments, such as when prices are sticky, irrespective of whether bonds and credit are perfect substitutes or not. But if the economy has imperfect price adjustment and if bonds and credit are not perfect substitutes, the B&B model implies different effects of monetary policy on aggregate demand, and consequently on output, from those in the IS–LM model. This also applies to our model.

Given our assumption for comparative static analysis that $\pi^{\mathrm{e}} = 0$, $R = r$ and $R^{\mathrm{L}} = r^{\mathrm{L}}$, the four endogenous variables in these three equations are y, P, r and r^{L}. Combining the IS, LM and CC equations to eliminate r and r^{L}, we end up with the aggregate demand equation:

$$y^{\mathrm{d}} = y^{\mathrm{d}}(\underset{-}{P}; m, g, \rho) \tag{19}$$

where m is a monetary policy variable (money supply and/or interest rate) and g is a fiscal policy one. The commodity supply side of our model has a production sector in which credit and bonds play different roles: changes in credit change the working capital of firms. Its earlier analysis gave the commodity supply function as:

$$y^{\mathrm{s}} = y^{\mathrm{s}}(\underset{-}{r^{\mathrm{L}}}, \rho) \tag{20}$$

37 The assessment of the literature on this point by Kashyap and Stein (1994) is that loans and bonds are not perfect substitutes.

38 For this assessment of the empirical information, see Kashyap and Stein (1994).

Therefore, equilibrium in the commodity markets yields the output equation as:

$$y = y^{d}(P; m, g, \rho) = y^{s}(r^{L}, \rho) \tag{21}$$

The credit market equilibrium condition derived earlier was:

$$P \cdot L^{d}(r^{L}, y) = L^{s}[\lambda(r^{L}, r)(1-\gamma)D, \rho] \tag{22}$$

The output equation (21), the CC condition (22) and the LM condition (13) can be combined by eliminating r and r^{L}. This yields the output produced as:

$$y = y((1-\gamma)D; g, \rho) \quad \partial y/\partial((1-\gamma)D) > 0 \tag{23}$$

Note that in our model the monetary policy variables determine the supply of loans, which enters as one of the components of working capital in the indirect production function. Hence, there is a positive relationship between y and $(1-\gamma)D$, so that an expansionary (contractionary) monetary policy increases (decreases) output.[39] Output also depends on ρ, our proxy for the riskiness and quantity of credit. If we interpret an increase in ρ as a loosening of credit rationing, y would depend positively on ρ.

16.5 Impact of monetary and fiscal policies

Impact of monetary policies on credit conditions

First, consider the impact of a tight monetary policy that increases both the bond and credit interest rates, thereby decreasing investment and the aggregate demand for commodities. Such an effect on aggregate demand also occurs in the standard IS–LM analysis. In addition, in our model, the fall in aggregate demand worsens business conditions for firms, thereby reducing their net worth and increasing the riskiness of credit, so that banks reduce the proportion of the portfolio that they wish to allocate to credit. This action will reduce the credit supply and further increase the credit interest rate. However, a sufficient shift of the banks' portfolio to bonds could end up increasing the banks' demand for bonds, so that the overall effect on the bond rate would become negative.[40] A smaller shift would still leave the bond rate higher, but moderate its increase. Therefore, in our money–bonds–credit (M–B–C) model, a tight monetary policy could lead to a smaller increase in the bond rate or a fall in it, while raising the credit interest rate. In any case, the spread between the credit interest rate and the bond rate will increase.

The impact of a contractionary monetary policy on short-run output proceeds through two channels. One is through the restriction in aggregate demand, which decreases output and prices if there are market imperfections of the type emphasized by the new Keynesians

39 The limit to the expansionary impact occurs if monetary policy increases the supply of working capital beyond its maximum demand as output increases.

40 Bernanke and Blinder (1988) cite March to July 1980 as a period during which a tight monetary policy in the USA reduced the government bond rate. Bernanke (1983) attributed the length of the Great Depression to a reduction in the supply of loans arising from the increase in their riskiness and banks' demand for increased liquidity/reserves due to an increase in the possibility of runs on them.

(see Chapter 15), but only decreases the price level if there are no market imperfections. The other channel of impact proceeds through the policy's impact on working capital: the contractionary monetary policy decreases the availability of credit and reduces the working capital of firms, which reduces their output. Therefore, the contractionary monetary policy reduces short-run output whether there are market imperfections or not, but the reduction in output is likely to be greater if there are market imperfections.

An expansionary monetary policy will improve business conditions by increasing aggregate demand and the sales made by firms. This will make credit to firms less risky and tilt the banks' portfolio composition towards credit from bonds. The expansionary monetary policy will lower both the credit and bond rates, while the portfolio composition shift will lower the credit interest rate further and moderate the fall in the bond rate. These will increase aggregate demand in the economy and would decrease output under a new Keynesian context of market imperfections. In addition, the greater availability of credit will increase the working capital of firms and, unless they were already at the maximum level needed for the current level of production, would increase output.

Impact of fiscal policies

An expansionary fiscal policy, with deficits financed by bond issues, will increase aggregate demand in the usual manner in the IS–LM and AD–AS analysis. This will increase output if there are market imperfections. The expansionary fiscal policy will also raise the bond rate, thereby causing a readjustment of banks' portfolio such as to decrease their credit supply. This has two effects. One is to raise the credit interest rate; the portfolio shift to credit from bonds will further increase the bond rate while moderating the increase in the credit interest rate. The second effect of the decrease in the supply of credit, and of the increase in the credit interest rate, is to reduce the working capital of firms, which will reduce output. Hence the expansionary fiscal policy has two contrary effects: in the presence of market imperfections, the increase in output due to the increase in aggregate demand, and the fall in output because of the decrease in working capital. The net effect is indeterminate when there are market imperfections. However, in the absence of such imperfections, the net effect will be negative: the expansionary fiscal policy will decrease output. The impact of a contractionary fiscal policy will be the opposite.

Impact of other causes of variations in credit supply or demand

Suppose that the riskiness of credit increases because of a negative shock, as would occur in a downturn in the economy, and adversely affects firms' ability to repay credit provided to them. This would reduce banks' supply of credit, so that, given bank deposits, the banks' demand for bonds would rise.[41] While the decreased supply of credit will raise the credit interest rate, the increased demand for bonds will lower the bond rate. The latter, as in the IS–LM model, will increase aggregate demand, which could, if there are new Keynesian market imperfections, increase output. However, the decrease in credit will reduce firms' working capital and output. Hence the net effect on output is indeterminate in the presence of market imperfections, but would be negative otherwise.

41 The banks' demand for free reserves may also rise, though, in many modern economies, changes in free reserves tend to be of minor significance.

A negative productivity shock would reduce productivity and demand for both labor and working capital. It also affects firms' ability to repay their loans and their borrowing through short-term corporate paper, so that it increases their risk. Therefore, a productivity shock would decrease output more in the short-run than in the long-run because of its impact on the short-run supply of credit.

Relative size of impact on small versus large firms

Our model relates credit supply and the credit interest rate to the working (and possibly physical) capital of firms and thereby to their short-run production/output of commodities. On this basis, assume that a decrease in the supply of credit reduces the working capital of the borrowing "small" firms and their production more than it does that of relatively large firms which have prearranged their short-run working capital needs through issues of bonds and retained earnings at the beginning of the current short-run. Further, there would be a flight to quality (less risky loans) as credit supply decreases, so that there would be differential effects; the effect will be more severe on the more credit-dependent small and intermediate-sized firms.

If the number of firms that rely on credit to facilitate production is significant enough in the economy, a decrease in the supply of credit will produce a fall in the output of the borrowing firms and, therefore, of the economy. A corresponding decrease in the demand for bonds will not elicit a similar impact in the short-term since this would not reduce funds already raised through bond issues and so will not reduce the working capital of firms. It may, however, affect the ability and cost of raising funds for the future through subsequent bond issues. Therefore, in the short-run, the impact on production and employment would be greater and occur faster for a decrease in the credit interest rate and/or the supply of credit than of a corresponding fall in the bond rate or decrease in the demand for bonds.

16.6 Instability in the money and credit markets and monetary policy

Controllability of credit supply by the central bank

Assuming that the credit market is an important independent source of effects on real output, the relevant question from the perspective of central bank policy is whether or not it can control the demand or supply of credit in some manner or control the credit interest rate. Since credit is fungible between bank loans, trade credit and short-term commercial paper, and trade credit has a certain amount of elasticity, a mild restrictive monetary policy that reduces bank loans somewhat may just be offset by an extension of trade credit, without any impact on working capital. This would eliminate the direct impact of the contractionary policy on output supply. Consequently, only the demand channel of monetary policy effects would be relevant. By comparison, a contractionary monetary policy severe enough to reduce working capital overall would have effects on both the supply side and the demand side. This implies a non-linear impact of monetary policy on output. The converse should roughly apply for the effects of expansionary monetary policies.

Central banks in several countries, such as the USA and Canada, pursue monetary policy through changes in the monetary base or/and the bond interest rate but do not pursue policies that directly affect the demand or supply of credit or the credit interest rate. Indirect central bank control of the supply of credit by banks through changes in the monetary base, brought about by open market operations, occurs through the latter's impact on bank deposits.

To the extent that the impact of a decrease in the monetary base on bank deposits is neutralized by banks through borrowing in the bond markets, such as by the sale of the banks' own short-term bond issues (such as certificates of deposit in the USA) to the public, the central bank's control over the loan supply by banks through changes in the monetary base will be diluted.

General monetary policies are likely to differentially affect both the credit and the bond markets in a manner determined by the economy and the lenders' portfolio decisions. They cannot be selectively targeted to the credit market alone. Therefore, their pursuit would be a somewhat undiscriminating instrument for manipulating the credit supply and/or its interest rate alone. However, central banks in many countries do pursue distinct policies affecting the credit market. Among such policies are selective ones that set the interest rates on credit or specify the allocation of banks' portfolio, or order specified increases in credit to specific sectors such as agriculture, exports and housing. To illustrate, on the purchase of durable consumer goods financed through installment credit,[42] some central banks set the down/initial payment and the period over which the loan has to be repaid.

Short-run versus long-run effects in Keynesian and neoclassical models

The preceding analysis has been short-run. In this short-run, changes in credit change firms' working capital, which alters output and employment. They also change aggregate demand through their impact on the credit and bond rates and, through them, on investment and consumption. If the economy is Keynesian, with imperfectly competitive firms setting prices and with some form of price rigidity (e.g. due to menu costs under a Calvo-type price adjustment mechanism), a change is demand produces a change in both output and prices (Clarida *et al*., 1999). Therefore, the Keynesian economy has two mechanisms, which together produce a change in output following a change in supply of loans. But if the economy is neoclassical, with perfect competition in all markets, changes in aggregate demand bring about a change only in the price level, so that a change in credit supply would alter output only through its impact on working capital. Therefore, in the short-run, a restrictive monetary policy would produce a decline in output in both the neoclassical and Keynesian models, but with a greater decline in the Keynesian case.

In the long run, changes in credit are only one mode of adjusting working capital, since retained earnings and bonds can also be used for this purpose. The economy will then have the amount of working capital required for output at the full-employment level, and changes in credit will not affect output. Further, the short-run price and nominal wage rigidity adduced by Keynesians will not exist in the long-run, so that changes in aggregate demand will not cause output to differ from its full-employment level. Hence, there is long-run neutrality of loans, just as of the money supply, whether the model is neoclassical or Keynesian and irrespective of whether a distinction is made between credit and bonds.

The financial instability hypothesis

Hyman Minsky (1986), along with other economists in the post-Keynesian tradition, argues that the financial sector is inherently unstable and possesses the capacity for destabilizing the real economy. The following provides some flavor of these arguments. As a boom progresses, firms increase their investment spending. To finance it, they increase their debt relative to their

42 Such policies go under the name of hire-purchase or installment-credit controls.

income flows, with their debt instruments becoming increasingly speculative. The financial markets are willing to absorb these during the boom as firms' profits rise and the financial markets, in an atmosphere of euphoria/exuberance and of "easy money," assess the risk of bond holdings to be relatively low. This euphoria also reduces the degree of risk aversion of lenders. The resulting disregard for risk leads to an expansion of credit, some of it extremely risky, which makes the financial sector vulnerable to a variety of shocks,[43] some of which can trigger a financial crisis. During this crisis, there occurs a reassessment of the riskiness of bonds and credit, as well as increasing risk aversion, so that the supply of funds to the credit market falls and their cost rises. The consequence is tighter credit, with access to funds becoming closed to some firms and reduced for others. This forces firms that rely on such funds for working capital and for financing short-term investment, e.g. in inventories, to cut back on them. The crisis may also be accompanied by cutbacks in consumer expenditure. These bring about reductions in production and investment in the economy as a whole. The resulting fall in profits intensifies the trend toward tighter credit and rising interest costs. The boom-time euphoria gets replaced by pessimism and panic during the downward spiral.

This story can be embellished by that on the existence of herds (Chari and Kehoe, 2002) in financial markets. In the absence of hard information on the future profitability or solvency of borrowers, in some cases, though not in others, an increasing perception of risk by some lenders is followed by a similar reassessment of risk by other lenders, so that the funds made available to borrowers as a whole decrease and trigger a decrease in production. Conversely, in some cases, though not in others, a decreasing perception of risk and aggressive lending by some financial institutions evokes a similar response from others, causing a general stampede toward easier credit, resulting in production increases. Business and consumer confidence on the future course of the economy, aggregate demand and job availability, are also quite susceptible to the herd phenomenon.

16.7 Credit channel when the bond interest rate is the exogenous monetary policy instrument

Chapter 13 argued that the central bank might use the interest rate as its primary instrument of monetary policy and might follow a Taylor rule for this purpose. This changes the preceding analysis by making the supply of money endogenous. In this context, suppose that the central bank pursues an expansionary monetary policy by lowering the bond interest rate. This has two immediate effects. One, to ensure equilibrium in the financial markets, the central bank has to ensure an adequate money supply by increasing the monetary base. Two, the decline in the bond rate makes credit more attractive in banks' portfolios, which causes banks to increase their supply of credit for a given money supply. This substitution of credit for bonds lowers the credit interest rate, which induces firms to take more loans. Hence, the expansionary monetary policy increases credit in the economy, which increases working capital. Therefore, the effects of an expansionary monetary policy implemented through a decrease in the bond rate are similar to those analyzed above for an expansion of the money supply.

43 Examples of such shocks include interest rate increases by the central bank in order to fight inflationary pressures, unexpected default by some borrowers, lower realized profits by firms than were expected, etc.

However, note that the impact on working capital is more directly through the money supply and less through interest rates.

16.8 The informal financial sector and financial underdevelopment

All economies also have an informal financial sector. Borrowers in this market usually do not have access to credit from the organized financial institutions such as banks, bond and stock markets, etc., but have to borrow from "moneylenders." This informal financial sector shares many of the characteristics of the "organized loan market" mainly operated by banks: normally, a particular borrower has access to credit from one or a few lenders, the credit is based on personal knowledge of the borrower and can often be recalled on demand or short notice. Therefore, for our analysis, the loan market can be taken to encompass credit given in both the organized and the informal financial sectors. Adding in the latter makes the credit market much more significant in economies with a substantial informal financial sector.

The existence of credit markets with quantity rationing means that the amount of credit extended is likely to be more closely related to the money supply than the credit interest rate is to the bond rate. This is especially so in financially underdeveloped economies in which the interest rates charged in the informal financial sector are likely to be only loosely related to the bond rates in the organized financial markets. In this case, the best monetary policy instrument is more likely to be the money supply rather than the interest rate. Consequently, even if the interest rate proves to be the more appropriate instrument in developed economies, as some estimations of the Taylor rule show (see Chapter 13), it need not be so for the developing economies.

16.9 Bank runs and credit crises

The financial system is prone to several special characteristics. Among these is contagion and runs, which create "sympathetic co-movements" in asset prices and liquidity.

Contagion is the spread of similar sentiments across financial institutions and instruments. For example, at the level of institutions, a widespread belief that a bank is about to fail leads to suspicions of the vulnerability of some other banks. At the level of assets, for example, a decline in the share prices of a firm in a particular industry leads to declines in the share prices of other firms. Contagion, therefore, leads to sympathetic co-movements across the institutions and assets of the financial system, and can be empirically observed in the self-reinforcing character of declines in asset prices. They also lead to a "herd movement" in which traders and individual investors rush to avoid losses by selling their asset holdings.

A run occurs because financial institutions typically hold assets with a longer maturity than their liabilities. In the case of banks, banks' liabilities are mainly withdrawal of deposits on demand, while their assets mainly consist of short-term bonds, loans, mortgages, etc. If there is sudden widespread withdrawal of deposits (a run), the bank's assets cannot be liquidated soon enough to meet the withdrawals, thereby forcing losses if the assets were liquidated at short notice, or insolvency for the bank. In addition, contagion can spread a run on one bank to other banks. Similar problems can afflict other financial institutions.

Financial systems have evolved several mechanisms to counter runs and contagion. Among these are the markets for trading reserves on an overnight basis (the federal funds market in the USA), the lender-of-last-resort function of central banks (see Chapter 11), insurance of bank deposits, and banking supervision to ensure proper financial practices. However, while

these reduce the possibility of runs, runs can and still do occur even in financially developed economies.

The relevance for the credit channel of runs on financial institutions and contagion among them is that their occurrence affects the availability and cost of credit in the economy, which can have real effects. Consequently, one of the roles of the central bank is to prevent their occurrence and, if they do hit, mitigate their effects on the financial system and the economy.

16.10 Empirical findings

As discussed in Chapter 2, monetary policy can affect aggregate demand through a number of transmission channels or mechanisms. Of these, the direct channel operates through the spending of excess money balances directly on commodities, while the indirect channel operates through a change in the interest rate, which alters investment. The indirect transmission channel is embodied in the IS–LM framework. The impact of money on output through these two channels has been labeled by some authors the "money view," as against the credit channel. The open economy also allows monetary policy to affect the balance of payments, changes which impact on domestic expenditure and output.

Among studies that did not find a significant independent impact of the loan channel on output are those of Oliner and Rudebusch (1996) and Driscoll (2004). Driscoll uses panel data from states within the USA to examine the impact of bank lending on output. He finds that shifts in money demand have large and statistically significant effects on the supply of bank loans, so that monetary policy can affect the latter. However, bank loans only have small, statistically insignificant, impact on output. Therefore, the bank lending channel is not a significant, independent contributor to the impact of monetary policy on output. Note that most such findings relate to the impact of bank loans, not of credit as a whole, on output. This chapter implies that changes in the overall credit supply affect output. If shifts in bank loans were offset by responsive shifts in other forms of credit, which are trade credit and short-term commercial paper, shifts in bank loans would not affect output.

Further, the regression of output on the money supply yields an estimate of the total impact of shifts in it on output. Since money supply and credit often tend to move together, in a regression of output on money the regression coefficient of money would tend to include within it at least some of the impact of changes in credit on output. If the estimation is then extended to include credit, in addition to money, as an explanatory variable, the marginal increase in predictive power may not prove to be significant. This is especially so for bank credit since bank credit and bank deposits move together because they reflect different sides of the banks' balance sheet. This makes it difficult to find a separate impact of bank loans in addition to that of money (King, 1986; Romer and Romer, 1990; Ramey, 1993; Walsh, 2003, Ch. 7). However, as the asset-backed security crisis in the USA in 2007 illustrates, shifts in the riskiness of credit and its perception and in the supply of credit can have distinct effects on output. Significant shifts of this kind tend to occur rarely, so that their impact may be more visible in detailed case studies of episodes in which the movement of credit diverged from that of money supply.

On this issue, Bernanke and Blinder (1988) compared the correlations between the growth rates of (nominal and real) GNP and money with those between GNP and credit[44] for the USA.

44 The money measure used was M1. Credit was "the sum of intermediated borrowing by households and businesses" from Flow of Funds data.

The correlations of GNP with money were higher for 1953:1–1973:4 than for credit but lower for 1979:4–1985:4. They also used simple least squares to estimate money and credit demand functions incorporating a partial adjustment model. While they could not check for parameter instability, the variance of the residual of the money demand equation was smaller for 1974:1–1979:3, but larger for 1979:4–1985:4, than that of the credit demand equation. They concluded that money demand shocks became relatively greater in the 1980s. While such evidence is suggestive rather than compelling and does not establish the direction of causality, Bernanke and Blinder claim that it may now be better for central banks to target credit rather than money.

Among studies supporting a distinctive lending channel, Kashyap and Stein (1993, 2000)[45] provide evidence supporting the distinction between loans and bonds and show that neither banks nor firms were indifferent between them. Consequently, changes in the loan interest rate or the loan supply have a different impact on aggregate demand than a corresponding change in the bond rate and the demand for bonds. In addition, this chapter's analysis views credit as providing part of the working capital needs of small and intermediate firms. A decrease in the availability of credit relative to their demand forces a decrease in their working capital and in their production, so that credit has a distinctive impact on both aggregate demand and supply in the economy. Bernanke (1986) finds that lending shocks have a sizeable effect on aggregate demand. Bernanke and Blinder (1992) report that as banks adjust their credit in response to monetary shocks, the decline in credit reduces output growth. Gertler and Gilchrist (1994), among many other studies, report that monetary tightening, such as by an increase in the interest rate, by the Fed often leads to a decline in bank loans to small firms and that the inventory investment of small firms is especially sensitive to such changes.

However, even if credit is an endogenous variable and changes in it are not sufficiently exogenous to changes in money or the bond rate, the credit structure of the economy shapes the dynamic response of real economic activity to monetary policy, so that it could still play an important role in the way that the impact of financial disturbances is propagated in the economy. Intuition indicates that the composition of spending among industries, firms and consumers does depend on the credit and bond structure of the economy. The compositional relevance of the credit channel can be established by examining shifts in the composition of economic activity among industries and firms in response to monetary policy shocks, and is supported by several empirical studies (see Kashyap *et al*., 1993; Gertler and Gilchrist, 1994; Lang and Nakamura, 1995; Ludvigson, 1998). To illustrate, Ludvigson reports that a tight monetary policy reduces bank consumer loans, which reduces real consumption expenditures, so that the composition of aggregate expenditures changes.[46]

Since different economies, especially financially developed as opposed to undeveloped economies, can have different bond and credit structures, international comparisons of the overall impact of monetary policy on output may provide some, though indirect, evidence on the relative importance of the credit channel. More direct evidence would have to come from the compositional effects of monetary policy among industries, firms and consumers.

45 Kashyap and Stein (2000) demonstrate that lending by small banks is relatively more sensitive to monetary policy. However, Ashcraft (2006) and Ashcraft and Campello (2007) report that the *average* impact of monetary policy on lending by banks is not statistically significant, while changes in the creditworthiness of borrowers, and therefore the balance sheet channel, do affect the response of bank lending to monetary policy.

46 Note that the empirical evidence of the compositional effects does not provide much guidance on the magnitude of the *quantitative* importance of credit effects in the overall impact of monetary policy on output.

These effects should differ among countries and be most visible in comparisons between the financially developed and the undeveloped economies.

Conclusions

The extensive consideration of market imperfections in macroeconomic modeling in recent years has extended to their role in financial markets. The imperfections in credit markets arise from adverse selection, moral hazard and monitoring and agency costs, and financial underdevelopment. Consequently, bonds and credit, which includes loans, are not perfect substitutes, so that their different characteristics can be better accommodated by classifying financial assets into three categories, money, bonds and credit, rather than two, money and bonds, which is the pattern of the IS–LM and IS–IRT models. The addition of information imperfections and credit help to explain many of the distinctive features of financial markets and their impact on the real economy.

This chapter relies on the addition of two innovations to the IS–LM and IS–IRT models of aggregate demand. One of these is to use information imperfections to draw a distinction between bonds (including equities) and credit. The other is the use of an indirect production function, which has working capital as an input since it facilitates the purchases of inputs (labor, raw materials and intermediate goods) bought prior to production and sales. Working capital usually comes from retained earnings, bonds (including stocks) and credit (including loans). Of these, this chapter's model simplified the analysis by assuming that the amount of working capital obtained by the firm from retained earnings and bonds does not vary in the short-run, but can be varied by the firm in the long-run.

Expansionary monetary and fiscal policies have somewhat different effects on output in our model. An expansionary monetary policy increases credit supply and reduces the credit interest rate, as well as increasing aggregate demand. An expansionary fiscal policy increases output in a new Keynesian model with market imperfections, but not in a neoclassical one (without market imperfections). The monetary policy, in our model, but not in the new Keynesian or neoclassical ones, increases the amount of working capital and output. Therefore, the expansionary (contractionary) monetary policy causes an unambiguous increase (decrease) in output.

Imperfections in financial imperfections imply that the decrease in credit in financial panics or runs will add to the impact of any decrease in the money supply on output. This reasoning implies that the contractionary impact of the financial sector in the Great Depression of the 1930s occurred not only through the decrease in the money supply but also through much more intense credit rationing, which worsened the fall in output and extended its duration.

Note that our definition of loan suppliers included moneylenders among banks. Developing economies have a larger proportion of firms that rely on loans rather than bond issues for their working capital needs. They also have a relatively large informal financial sector. Our analysis shows that the impact on production of a fall in the supply of loans is likely to be more intense and speedier in developing economies than in financially developed ones.

Since a change in the money supply is positively related to changes in the amount and cost of credit, the credit channel intensifies the impact of monetary policy on aggregate demand beyond that encompassed in the "money view," i.e. monetary transmission effects on aggregate demand through the bond interest rate. Further, the overall effect of money on output now includes, in addition to effects through aggregate demand, a working capital

effect, since credit, but not money itself, is a component in the proposed short-run production function.

There are vertical layers among the institutions that provide credit. The link between money supply and these credit layers tends to be imprecise, so that the supply of credit cannot be controlled accurately enough by monetary policy. Hence the central bank may only be able to moderate, but not fully offset, the effects of a credit crisis, arising say from a heightened perception of the riskiness of lending to borrowers in credit markets, on output and employment, through an expansionary monetary policy. The credit channel, therefore, can dilute the central bank's control of the economy, especially during a credit crunch or credit euphoria.

Appendix A

Demand for working capital for a given production level in a simple stylized model

To illustrate the firm's demand for working capital and the differentiation between it and the firm's money holdings, we adapt Baumol's (1952) inventory analysis of the demand for money as a medium of payments in the presence of an alternative financial asset. The two financial assets in the following analysis are money and loans/credit. We assume that the firm purchases its input (labor, raw materials and intermediate inputs) and has to pay for them in an even stream during the period. These payments are made in advance of production and sales, with sales assumed to occur only at the end of the period. For simplification, the values of the purchases and sales revenues are assumed to be identical. The firm finances its purchases through a loan, arranged from a bank at the beginning of the period in the form of an overdraft. The loan to the firm is repaid at the end of the period from its sales revenue.

Let the total cost of inputs (and the sales revenue) equal $\$Y$. The firm is assumed to withdraw from its overdraft the amounts needed for its payments in an even stream through the period, with the amount drawn each time being $\$z$ and the number q of withdrawals equal to Y/z. Since the firm will withdraw z at the beginning of the period, after an interval equal to $1/q$ of the period, after another interval equal to $1/q$, and so on, the average amount withdrawn and spent on purchases is:

$$\begin{aligned} K^{\mathrm{w,d}} &= z + z(q-1)/q + z(q-2)/q + \cdots + z/q \\ &= (q+1)z/2 = \tfrac{1}{2}\,Y + \tfrac{1}{2}\,z \end{aligned} \tag{24}$$

Since $qz = Y$, the firm's average money balances will equal $M/2$, so that:

$$K^{\mathrm{w,d}} = \tfrac{1}{2}\,Y + M \tag{25}$$

Hence, the demand for working capital is greater than for money balances. In real terms, this demand becomes:

$$K^{\mathrm{w,d}}/P = \tfrac{1}{2}\,y + m \tag{26}$$

where $y = Y/P$ and $m = M/P$.

Equation (25) specifies the *maximum* amount of working capital needed by the firm to finance a given amount of purchases of inputs. The firm can economize on its working capital

by reducing its output or/and by diverting some of its labor to save on working capital. For a given maximum need for working capital, the firm may also be able to substitute to some extent trade credit for loans, but this will also usually involve some diversion of labor to make alternative trade credit arrangements with the suppliers and buyers from the firm. This possibility is investigated in Appendix B.

In the long run, the firm can provide for its working capital needs through bonds, retained earnings, trade credit and loans. Of these, in the short-run, credit is the only component that is taken to be variable in this chapter.

Appendix B

Indirect production function including working capital

The firm's demand for working capital is the average amount of the "funds" that it wants to hold in order to carry out its purchases of labor services and other inputs (raw materials and intermediate goods). The following analysis justifies the appearance of working capital as an input in the production function and the determination of its optimal amount through profit maximization.

Assume that the firm's output depends on its physical capital and the part of its employment that it uses directly as an input in production. However, it has to divert some of its workers to carrying out transactions involving purchases of inputs and the sale of its output. If the firm held only a small and relatively inadequate amount of working capital, it would have to employ workers in juggling its working capital to carry out the required transactions of purchase and sale of commodities, as well as payments to workers. Firms usually use a mix of trade credit and their own working capital.[47] However, arranging and using the former implies a relatively higher amount of labor time than if the firm holds adequate working capital. The use of working capital, therefore, allows the firm to economize on the workers it has to divert to making payments.

The preceding arguments yield the representative firm's production function as:

$$y = y(n_1) \quad \partial y/\partial n_1 > 0, \partial^2 y/\partial {n_1}^2 < 0 \tag{27}$$

where y is the firm's output and n_1 is the amount of labor directly involved in production. Total employment by the firm is $n_1 + n_2$, where n_2 is the labor used in the payments and receipts processes so that:

$$n_1 = n - n_2 \tag{28}$$

where n is total employment. For given n, $\partial n_1/\partial n_2 < 0$.

For the labor used in carrying out exchanges, and using the firm's output y as the proxy for the number of transactions involved in purchasing inputs, the use of n_2 workers is taken to be given by:

$$n_2 = n_2(k^{\mathrm{w}}, y) \tag{29}$$

47 Guariglia and Mateut (2006) report on the existence of substitutability between trade credit and credit (loans) at the micro level for the UK, and that this substitution weakens the loan channel.

where $k^{\mathrm{w}}(= K^{\mathrm{w}}/P)$ is the firm's real working capital, $\partial n_2/\partial k^{\mathrm{w}} \leq 0$ and $\partial n_2/\partial y > 0$. The specific form of $n_2(.)$ would depend on the trading and payments technology of the economy and would shift with that technology. Innovations in the financial system, such as the use of direct deposit of salaries into the workers' accounts and payments to suppliers by electronic transfers, would reduce the demand for real balances for transactions associated with a given level of output and shift the transactions technology function. It will also depend on the availability and the flexibility of trade credit.

From (27) to (29),

$$\frac{\partial y}{\partial k^{\mathrm{w}}} = \frac{\partial y}{\partial n_1}\frac{\partial n_1}{\partial n_2}\frac{\partial n_2}{\partial k^{\mathrm{w}}} \geq 0$$

A specific form of (29) is the proportional expression:

$$n_2/y = \phi(k^{\mathrm{w}}/y) \tag{30}$$

where $\phi' = \partial\phi/\partial(k^{\mathrm{w}}/y) \leq 0$. For this function, the firm reaches "saturation" in real balances relative to its output when $\phi' = 0$, which will occur when the firm holds the maximum amount of working capital, as derived in Appendix A. From (27) to (30),

$$y = y(n - y \cdot \phi(k^{\mathrm{w}}/y)) \tag{31}$$

which can be rewritten as the indirect production function:

$$y = f(n, k^{\mathrm{w}}) \tag{32}$$

where $\partial y/\partial k^{\mathrm{w}} \geq 0$. Hence, the use of working capital by the firm increases its output, with its marginal product being positive up to the saturation point $k^{\mathrm{w,max}}$. Up to this point, the use of working capital allows the firm to reduce the labor allocated to payments, thereby increasing the labor allocated directly to production, which increases the firm's output for a given amount of employment.

Profit maximization and the optimal demand for working capital

The firm is assumed to operate in perfect competition in all (output and input) markets and to maximize profits. Its profits are given by:

$$\Pi = PF(n, k^{\mathrm{w}}) - Wn - R^{\mathrm{L}} \cdot Pk^{\mathrm{w}} - F_0 \tag{33}$$

where Π is profits, P is the price level and F_0 is fixed cost, which includes the firm's commitment to pay interest on its existing bonds. n is employment, W is the nominal wage rate and k^{w} is the amount used of real working capital.

The first-order conditions for maximizing profits with respect to n and k^{w}, are:

$$P \cdot \partial F/\partial n - W = 0 \tag{34}$$

$$P \cdot \partial F/\partial k^{\mathrm{w}} - P \cdot R^{\mathrm{L}} = 0 \tag{35}$$

Solving (34) and (35) yields the demand functions:

$$n^{\mathrm{d}} = n^{\mathrm{d}}(w, R^{\mathrm{L}}) \tag{36}$$

$$k^{\mathrm{w,d}} = k^{\mathrm{w,d}}(w, R^{\mathrm{L}}) \tag{37}$$

where $w = W/P$. The supply of output is given by substituting (36) and (37) in the indirect production function (32). Its functional form is:

$$y = f(w, R^{\mathrm{L}}) \tag{38}$$

With $\pi^{\mathrm{e}} = 0$, this function becomes $y = f(w, R^{\mathrm{L}})$.

Demand for credit

The short-run nominal demand for credit is given by:

$$L^{\mathrm{d}} = Pk^{\mathrm{w,d}}(w, R^{\mathrm{L}}) - K^{\mathrm{w\#}} \tag{39}$$

where $K^{\mathrm{w\#}}$ is the nominal amount of working capital prearranged through bonds and retained earnings. Short-run variations in $Pk^{\mathrm{w,d}}$, therefore, produce corresponding variations in the short-run demand for loans. Hence:

$$L^{\mathrm{d}}/P = \psi(w, R^{\mathrm{L}}; K^{\mathrm{w\#}}) \tag{40}$$

Summary of critical conclusions

- Adverse selection, moral hazard and monitoring and agency costs imply quantity credit rationing, in addition to rationing by the interest rate charged.
- Imperfections in financial markets lead to a distinction between bonds and credit.
- Firms need working capital to facilitate the purchases of inputs and sale of output, so that it becomes an argument of their indirect production function.
- The assumption that the funds raised through bonds (including stocks) are given in the short-run, while those raised through credit (especially trade credit and bank loans) are variable, allows monetary policy to change the working capital of firms and its cost, so that it has a direct impact on the supply of commodities in the economy.
- The credit channel is one of several channels that determine the magnitude and lags in the impact of monetary policy on output.
- It is difficult to empirically estimate the marginal impact of the lending channel on the economy's relationship between money and output. However, its impact can be established convincingly in terms of its effects on the composition of expenditures among small and large firms, and between firms and households, and of output among industries.
- The credit channel is likely to be relatively more important in financially underdeveloped economies than in financially developed ones.

Review and discussion questions

1. Discuss the main characteristics of money, bonds, credit and equities/stocks in actual financial markets. What is gained and what is lost by having a macroeconomic model

with only two financial assets, money and bonds, with the latter including credit and equities?

2. Discuss the sources of imperfections in credit markets and their role in quantity and interest rate rationing.
3. Specify a model of aggregate demand with three financial assets, money, bonds and credit.
4. Given your model of aggregate demand when there are three financial assets, money, bonds and credit, and the standard classical production function, would disturbances in the loan market cause changes in output and employment? Your answer should present the determination of output and employment in this model.
5. Given your model of aggregate demand when there are three financial assets, money, bonds and credit, show how the new Keynesian model allows shifts in the credit market to alter output and employment.
6. What is working capital? Justify its inclusion in a production function and show the impact of a decrease in working capital on the supply of short-run output and employment.
7. Given your model of aggregate demand with three financial assets (money, bonds and credit) and the indirect production function with working capital (but without the new Keynesian Phillips curve), how can monetary policy produce an increase in short-run output and employment? Can it do so in long-run output and employment? Discuss.
8. Given your model of aggregate demand with three financial assets (money, bonds and credit) and the indirect production function with working capital and the new Keynesian Phillips curve, what are the effects of a contractionary monetary policy on output and employment in the short-run? What are its effects in the long run? Discuss.
9. Discuss why the credit channel is likely to be more important in financially developing economies than in developed ones, and discuss its implications for the choice between the money supply and the interest rate as the appropriate monetary policy instrument.
10. Discuss the significance of the credit channel in changing aggregate demand and output. What limitations, if any, on this significance are imposed by the addition of the expectations-augmented Phillips equation? What limitations, if any, on this significance are imposed by the addition of short-run money neutrality?

References

Akerlof, G. "The market for 'lemons': quality uncertainty and the market mechanism." *Quarterly Journal of Economics*, 84, 1970, pp. 488–500.

Ashcraft, A.B. "New evidence on the lending channel." *Journal of Money, Credit and Banking*, 38, 2006, pp. 751–75.

Ashcraft, A.B. and Campello, M. "Firm balance sheets and monetary policy transmission." *Journal of Monetary Economics*, 54, 2007, pp. 1515–28.

Baumol, W.J. "The transactions demand for cash: an inventory theoretic approach." *Quarterly Journal of Economics*, 66, 1952, pp. 545–56.

Bernanke, B.S. "Nonmonetary effects of the finanical crisis in the propagtion of the Great Depression." *American Economic Review*, 73, 1983, pp. 257–76.

Bernanke, B.S. "Alternative explanations of the money–income correlation." *Carnegie-Rochester Conference Series on Public Policy*, 25, 1986, pp. 49–99.

Bernanke, B.S. "Credit in the macroeconomy." *Federal Reserve Bank of New York Quarterly Review*, 18, 1992–93, pp. 50–70.

Bernanke, B.S. and Blinder, A.S. "Credit, money, and aggregate demand". *American Economic Review Papers and Proceedings*, 78, 1988, pp. 435–39.

Bernanke, B.S. and Blinder, A.S. "The federal funds rate and the channels of monetary transmission." *Amercian Economic Review*, 82, 1992, pp. 901–21.

Bernanke, B.S. and Gertler, M. "Agency costs, net worth, and business fluctuations." *American Economic Review*, 79, 1989, pp. 14–31.

Bernanke, B.S., Gertler, M. and Gilchrist, S. "The financial accelerator in a quantitative business cycle framework." In J.B. Taylor and M. Woodford, eds, *Handbook of Macroeconomics*, vol. 1C. Amsterdam: Elsevier North-Holland, 1999, pp. 1341–93.

Chari, V.V. and Kehoe, P.J. "On the robustness of herds." *Federal Reserve Bank of Minneapolis Working Paper* no. 622, 2002.

Clarida, R., Gali, J. and Gertler, M. "The science of monetary policy: a new Keynesian perspective." *Journal of Economic Literature*, 37, 1999, pp. 1661–707.

Driscoll, J.C. "Does bank lending affect output? Evidence from the U.S. states." *Journal of Monetary Economics*, 51, 2004, pp. 451–71.

Fama, E. "Banking in the theory of finance." *Journal of Monetary Economics*, 27, 1980, pp. 39–57.

Gertler, M. and Gilchrist, S. "Monetary policy, business cycles and the behavior of small manufacturing firms." *Quarterly Journal of Economics*, 109, 1994, pp. 309–40.

Guariglia, A. and Mateut, S. "Credit channel, trade credit channel, and inventory investment: evidence from a panel of UK firms." *Journal of Banking and Finance*, 30, 2006, pp. 2835–56.

Hubbard, R.G. "Is there a credit channel for monetary policy?" *Federal Reserve Bank of St. Louis Review*, 77, 1995, pp. 63–77.

Jaffee, D. and Russell, T. "Imperfect information, uncertainty and credit rationing." *Quarterly Journal of Economics*, 90, Nov. 1976, pp. 651–66.

Jaffee, D. and Stiglitz, J.E. "Credit rationing." In B. Friedman and F. Hahn, eds, *The Handbook of Monetary Economics*, Vol. II. Amsterdam: North-Holland, 1990, pp. 837–88.

Kashyap, A.K. and Stein, J.C. "Monetary policy and bank lending." In N.G. Mankiw, ed., *Monetary Policy*. Chicago: University of Chicago Press, 1994, pp. 221–56.

Kashyap, A.K. and Stein, J.C. "The role of banks in monetary policy: A survey with implications for the European monetary union." *FRB of Chicago Economic Perspectives*, 21, 1997.

Kashyap, A.K. and Stein, J.C. "What do one million observations on banks have to say about the transmission of monetary policy?" *American Economic Review*, 90, 2000, pp. 407–28.

Kashyap, A.K., Stein, J.C. and Wilcox, D.W. "Monetary policy and credit conditions: evidence from the composition of external finance." *American Economic Review*, 83, 1993, pp. 78–98.

Keynes, J.M. "Alternative theories of the rate of interest." *Economic Journal*, 47, 1937, pp. 241–52.

King, S.R. "Monetary transmission: through bank lending or bank liabilities." *Journal of Money, Credit and Banking*, 18, 1986, pp. 290–303.

Kiyotaki, N. and Moore, J. "Credit cycles." *Journal of Political Economy*, 105, 1997, pp. 211–48.

Lang, W.W. and Nakamura, L.I. " 'Flight to quality' in bank lending and economic activity." *Journal of Monetary Economics*, 36, 1995, pp. 145–64.

Ludvigson, S. "The channel of monetary transmission to demand: evidence from the market for automobile credit." *Journal of Money, Credit and Banking,* 30, 1998, pp. 365–83.

Minsky, H.P. *Stabilising an Unstable Economy*. New Haven, CT: Yale University Press, 1986.

Modigliani, F. and Miller, M.H. "The cost of capital, corporate finance, and the theory of investment." *American Economic Review*, 48, 1958, pp. 261–97.

Oliner, S.D. and Rudebusch, G.D. "Is there a broad credit channel for monetary policy?" *Federal Reserve Bank of San Francisco Economic Review*, 1, 1996, pp. 300–09.

Ramey, V. "How important is the credit channel in the transmission of monetary policy?" *Carnegie-Rochester Conference Series on Public Policy*, 39, 1993, pp. 1–45.

Romer, C.D. and Romer, D.H. "New evidence on the monetary transmission mechanism." *Brookings Papers on Economic Activity*, 1, 1990, pp. 149–98.

Stiglitz, J. and Weiss, A. "Credit rationing in models with imperfect information." *American Economic Review*, 71, 1981, pp. 393–410.

Walsh, C.E. *Monetary Theory and Policy*, 2nd edn. Cambridge, MA: MIT Press, 2003.

Williamson, S.D. "Costly monitoring, loan contracts and equilibrium credit rationing." *Quarterly Journal of Economics*, 102, 1987, pp. 135–45.

17 Macro models and perspectives on the neutrality of money

This chapter continues the discussion of the effectiveness of monetary policy. It starts with the compact Lucas–Sargent–Wallace model, which is a popular platform for the modern classical approach. It then examines popular Keynesian and new Keynesian compact models. The emphasis of this chapter is on the presentations of compact, testable models and their findings.

Economists' and central bankers' thinking on the relevance and impact of monetary policy is presented in the conclusions.

Key concepts introduced in this chapter

- Monetary policy ineffectiveness proposition
- Neutrality of money
- Lucas–Sargent–Wallace model
- Lucas critique of estimated equations
- Keynesian supply rule
- New Keynesian Taylor rule
- New Keynesian IS equation
- New Keynesian model
- Hysteresis

The use of compact models to examine the impact of monetary policy on the macroeconomy is common in the macroeconomic literature. This chapter examines several of these models. For the modern classical approach, we have selected the Sargent and Wallace (1976) model. For the Keynesian and new Keynesian approaches, we have selected the models of Gali (1992), Clarida *et al*. (1999) and Levin *et al*. (2001) for the closed economy, and that of Ball (1999, 2000) for the open economy.

Sections 17.1 and 17.2 analyze the effects of systematic and unanticipated money supply changes on output and prices in the context of the Lucas–Sargent–Wallace model. Section 17.3 shows the validity of the Lucas critique in this model. Sections 17.4 to 17.7 cover the empirical evidence on the issue of monetary neutrality. Sections 17.8 and 17.9 present compact forms of the new Keynesian models and examine their validity. Section 17.10 sums up the empirical evidence on money neutrality and Section 17.11 suggests getting away from dogma. Section 17.12 provides a brief comment on hysteresis in the overall context of long-run money neutrality.

The conclusions of this chapter present a summing-up of our knowledge as reflected in the writings of some of the major protagonists on the neutrality debate.

17.1 The Lucas–Sargent–Wallace (LSW) analysis of the classical paradigm

The Lucas (1972, 1973) model presented in Chapter 14 serves as the underlying supply behavior of the modern classical approach. Based on this supply rule, the 1976 Sargent–Wallace model and its variants represent the most commonly used format for deriving the implications of the modern classical model and testing them. Since this model is based on Lucas (1972, 1973) and incorporates the Lucas supply function, we refer to it as the Lucas–Sargent–Wallace (LSW) model. It explicitly specifies the markets for commodities and money, with the assumption of equilibrium in these markets. However, the labor market is not explicitly modeled but is replaced, in conjunction with the production function, by the Friedman–Lucas supply function.

For our presentation of this model, we specify the change in the supply of commodities by stating the expectational error-based Lucas supply function or rule[1] as:

$$\mathrm{D}y^{\mathrm{s}}{}_t = \alpha \mathrm{D}y_{t-1} + \beta(p_t - p^{\mathrm{e}}{}_t) + \mu_t \qquad 0 \leq \alpha \leq 1, \beta > 0 \tag{1}$$

where all lower-case variables are in *logs*, the superscripts s and d stand respectively for supply and demand, and:

y = output
y^{f} = full employment output
$\mathrm{D}y_t$ = output gap $(= y_t - y_t^{\mathrm{f}})$
p = price level
p^{e} = expected price level, with expectations formed one period earlier
μ = random term.

Note that Dy is the deviation from full-employment output and not the previous period's output. μ and the other random terms in this chapter have a zero expected value and are independent of the other variables in the model. Equation (1) differs from the Friedman–Lucas supply rule derived in Chapter 14 by including a lagged term in output, with $\alpha > 0$. This is done in order to capture the commonly observed serial correlation of output over time. This alteration can be explained by adducing adjustment costs for employment, so that the marginal product of labor depends on both the current and last period's output.[2] Note that (1) also differs from the Friedman–Lucas supply rule in that it allows an output gap to exist even when there are no errors in price expectations: current output can differ from its full-employment level if there was an output gap in the preceding period. However, the output gap gets eliminated at the rate α per period. (1) also allows the full-employment output to change between periods. As shown in Chapter 13, the dependence of $y^{\mathrm{s}}{}_t$ on $(p_t - p^{\mathrm{e}}{}_t)$ can be justified through the expectations-augmented Phillips curve, with nominal wages either fixed for the duration of the labor contract or fully flexible but with imperfect information about the price level and with expectational errors in perceived relative prices.[3]

1 We are subsuming the expectations-augmented Phillips curve in the Lucas supply rule in this chapter.

2 Such a term is needed for classical models since, in its absence, the deviations of output from full employment would depend only upon errors in expectations. Since the current classical models also assume rational expectations in which the errors in expectations are random, (1) with rational expectations would imply that output variations over time would be only random, even though business cycles show serial correlation in output.

3 Lucas (1973) specifies that production takes place on isolated points called islands; the selling price on the island is known but prices on other islands are not known. The demand for labor is a function of the real

The demand for output is specified as:

$$y^{d}_{t} = \theta(m_t - p_t) + \eta_t \qquad \theta > 0 \tag{2}$$

where:

y^{d} = aggregate demand
m = nominal money supply
η = random term.

Again, all variables are in logs. (2) represents the aggregate demand function and is a reduced-form relationship derived from the IS–LM relationships. It ignores fiscal policy for two reasons. One is that the effects of monetary and fiscal expansions in the IS–LM models are similar, so that keeping only one policy variable simplifies the model. In this sense, fiscal policy does influence aggregate demand and its stance is proxied in (2) through m_t. The other reason is that the new classical model with Barro's Ricardian equivalence theorem, outlined in Chapter 13 above, implies that fiscal deficits do not change aggregate demand. Only increases in the money supply do so, with the result that (2) becomes the accurate representation of the aggregate demand function, irrespective of the fiscal policy stance.

Equation (2) differs from the aggregate demand function in Lucas's model in Chapter 14 by making explicit the role of the nominal money supply in the determination of aggregate demand. But it also necessitates the specification of the money supply function, which is done by the monetary policy rule:

$$m_t = m_0 + \gamma \mathrm{D}y_{t-1} + \xi_t \qquad \gamma < 0 \tag{3}$$

Equation (3) makes the plausible assumption that the monetary authority increases the money supply if $\mathrm{D}y_{t-1} < 0$, i.e. if output last period was below the full employment level.

The equilibrium condition for the commodity market is:

$$y_t = y^{d}_{t} = y^{s}_{t} \tag{4}$$

which also translates to:

$$\mathrm{D}y_t = \mathrm{D}y^{d}_{t} = \mathrm{D}y^{s}_{t}$$

The above model has to be supplemented by an expectations hypothesis. Assuming the rational expectations hypothesis (REH), we have:

$$p^{e}_{t} = \mathrm{E}p_t \tag{5}$$

where $\mathrm{E}p_t$ represents the expected price conditional on information available at the beginning of period t and $(p_t - \mathrm{E}p_t)$ is a random variable with zero mean. In particular, p_{t-1} is part of this information set.

wage rate in terms of the island price, which on average in the aggregate for all islands equals the actual price, while the supply of labor is a function of the expected real wage in terms of the expected price level over all islands.

The complete LSW model, labeled in this chapter *LSW Model I*, consists of equations (1) to (5). With all variables in logs, this model is:

$$\mathrm{D}y^{\mathrm{s}}_t = \alpha \mathrm{D}y_{t-1} + \beta(p_t - p^{\mathrm{e}}_t) + \mu_t \qquad \alpha, \beta > 0 \tag{1}$$

$$y^{\mathrm{d}}_t = \theta(m_t - p_t) + \eta_t \qquad \theta > 0 \tag{2}$$

$$m_t = m_0 + \gamma \mathrm{D}y_{t-1} + \xi_t \qquad \gamma < 0 \tag{3}$$

$$y_t = y^{\mathrm{d}}_t = y^{\mathrm{s}}_t \tag{4}$$

$$p^{\mathrm{e}}_t = \mathrm{E}p_t \tag{5}$$

The basic question we wish to investigate within this model is whether monetary policy can be manipulated to increase output. More explicitly, we want to investigate whether there are particular values of m_0 and γ in (3) which optimize y_t. To answer this, we need to derive the reduced-form equation for y_t. From (1), (4) and (5), and taking the expectation of y_t, we have:

$$\begin{aligned} \mathrm{ED}y_t &= \alpha \mathrm{ED}y_{t-1} + \beta[\mathrm{E}p_t - \mathrm{E}(\mathrm{E}p_t)] + \mathrm{E}\mu_t \\ &= \alpha \mathrm{D}y_{t-1} \end{aligned} \tag{6}$$

Substituting (4) in (2) and taking its expectation, with $\mathrm{E}\eta_t = 0$, gives:

$$\begin{aligned} \mathrm{E}y_t &= \theta(\mathrm{E}m_t - \mathrm{E}p_t) + \mathrm{E}\eta_t \\ &= \theta(\mathrm{E}m_t - \mathrm{E}p_t) \end{aligned} \tag{7}$$

Subtracting (7) from (2) yields:

$$y_t - \mathrm{E}y_t = \theta(m_t - \mathrm{E}m_t) - \theta(p_t - \mathrm{E}p_t) + \eta_t \tag{8}$$

Subtracting y^{f}_t from both sides, we get:

$$\mathrm{D}y_t = \mathrm{ED}y_t + \theta(m_t - \mathrm{E}m_t) - \theta(p_t - \mathrm{E}p_t) + \eta_t \tag{8'}$$

where, from (3),

$$\mathrm{E}m_t = m_0 + \gamma \mathrm{D}y_{t-1} \tag{9}$$

so that:

$$m_t - \mathrm{E}m_t = \xi_t \tag{10}$$

From (1), (4) and (5),

$$p_t - \mathrm{E}p_t = (1/\beta)(\mathrm{D}y_t - \alpha \mathrm{D}y_{t-1}) - (1/\beta)\mu_t \tag{11}$$

In (8′), replacing the relevant terms on the right-hand side from (6), (10) and (11) gives:

$$\begin{aligned} \mathrm{D}y_t &= \alpha \mathrm{D}y_{t-1} + \frac{\theta\mu_t + \beta\theta\xi_t + \beta\eta_t}{\beta + \theta} \\ &= \alpha \mathrm{D}y_{t-1} + \Psi_t \end{aligned} \tag{12}$$

where:

$$\psi_t = \frac{\theta\mu_t + \beta\theta\xi_t + \beta\eta_t}{\beta + \theta}$$

Hence, the current output gap is the correction by α of last period's gap plus a new disturbance. (12) represents the core implication of the LSW model for the determination of output.

The ineffectiveness proposition of the LSW model for monetary policy

Since neither of the systematic policy parameters m_0 and γ occur in (12), the authorities cannot use systematic monetary policy to change y_t. Since ξ_t is in (12), errors in predicting money supply do affect y_t, but if the authorities were to increase such errors, it would only increase the variance of y_t without increasing the output level and, therefore, without constituting a sensible policy. There is therefore no optimal monetary policy in this model. Even though output can differ from its full-employment level, non-random policy does not have any long-run or even short-run effects on real output; it can neither cause nor improve nor worsen booms or recessions. These results are similar to those derived from the Lucas model in Chapter 14. The implication of the futility of systematic money supply changes to alter output – and by implication, employment and unemployment – is known as the "the ineffectiveness of demand policies" or the "demand policy irrelevance" result. This is so even if current output is below the full-employment rate. Hence, there is no stabilization role for monetary policy in this model. Note that, just as systematic money supply changes can alter aggregate demand but not output and employment, systematic exogenous increases in investment, consumption, net exports, etc., also cannot change output and employment by virtue of their impact on aggregate demand. But their random components can do so.

The LSW model does not support Keynesian policy recommendations on the use of monetary policy to reduce deviations from full-employment output. However, this is understandable since the model is not a Keynesian one to start with. In particular, its supply equation (1) is the Friedman–Lucas supply rule, which embodies, under rational expectations, the neutrality of systematic changes in the money supply.

Price level in the LSW model

The reduced form for p_t for this model is obtained by substituting (12)[4] in (2). This gives:

$$p_t = m_t - \frac{1}{\theta} y_t^{\mathrm{f}} - \frac{\alpha}{\theta} \mathrm{D}y_{t-1} - \frac{1}{\theta}\psi_t + \frac{1}{\theta}\eta_t \tag{13}$$

4 To do so, first replace $\mathrm{D}y_t$ by $(y_t - y_t^{\mathrm{f}})$.

Equation (13) implies that $\partial p_t / \partial m_t = 1$. Hence, since both p and m are in logs, prices rise proportionately with the *overall* increase in the nominal money supply, whether it is due to the systematic factors γ and m_0 or the random component ξ_t.

To find p^e_t, since $p^e_t = E(p_t)$, taking the rational expectation of (13) implies that:

$$p^e_t = \mathrm{E}(m_t) - (1/\theta)y^f_t - (\alpha/\theta)\mathrm{D}y_{t-1} \tag{14}$$

From (14) and (9),

$$p^e_t = m_0 - (1/\theta)y^f_t + \{\gamma - (\alpha/\theta)\}\mathrm{D}y_{t-1} \tag{15}$$

In (14) and (15), $\partial p^e_t / \partial \mathrm{E}m_t = 1$, so that the expected price level rises proportionately in the same period with the *systematic* component of the nominal money supply, and responds to the policy parameters γ and m_0, but not to the random part ξ_t of the money supply. Hence, increases in systematic money supply (and systematic demand) proportionately change both the price level *and* the expected price level but do not cause a change in output, whereas random changes in the money supply change the price level and output but not the expected price level.

Diagrammatic analysis of output and price level in the LSW model

The effect of a systematic monetary increase in the LSW model is shown in Figure 17.1. From (2), the aggregate demand curve AD has a negative slope and shifts to the right from AD_0 to AD_1 with a monetary expansion. From (1), the (short-run) aggregate supply curve SAS has a positive slope and shifts to the left from SAS_0 to SAS_1 with an increase in the *expected* price level. Since the latter, from (14), increases proportionately with a systematic monetary expansion, such a monetary expansion results in proportionate shifts of both curves, and the economy goes directly from point a to point c without an intermediate increase in output. However, a random increase in the money supply cannot be anticipated, so that the expected price level does not increase as a result and the supply curve does not shift from SAS. But the demand curve does go from AD_0 to AD_1, with the new equilibrium at point b causing an increase in both prices and output. Hence, while the systematic increases in money supply do not bring about an increase in output, unexpected changes in it do so.[5]

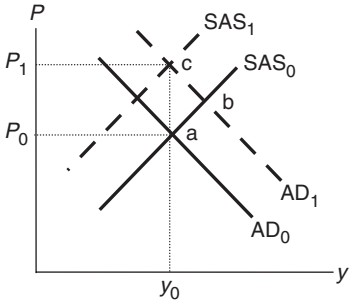


Figure 17.1

5 If we continue the story further, the unexpected increase in the current price level will have become anticipated in the following period, so that, barring new sources of shifts in aggregate demand and supply, the economy will

17.2 A compact (Model II) form of the LSW model

Since the price level is a function of the money supply and the expected price level is a function of the expected money supply, the above model with rational expectations is often replaced by the following more compact one, labeled by us as the LSW Model II.

Lucas supply rule:

$$\mathrm{D}y_t = \alpha \mathrm{D}y_{t-1} + \beta(m_t - m^{\mathrm{e}}{}_t) + \mu_t \qquad \alpha, \beta > 0 \tag{16}$$

Money supply rule:

$$m_t = m_0 + \gamma \mathrm{D}y_{t-1} + \xi_t \qquad \gamma < 0 \tag{17}$$

Rational expectations:

$$m^{\mathrm{e}}{}_t = \mathrm{E}m_t \tag{18}$$

Note that the assumption of equilibrium in the commodity market has already been incorporated in (16). From (17) and (18),

$$m_t - m^{\mathrm{e}}{}_t = \xi_t \tag{19}$$

Hence, from (16) and (19),

$$\mathrm{D}y_t = \alpha \mathrm{D}y_{t-1} + \beta\xi_t + \mu_t \tag{20}$$

where $\mathrm{D}y_t$ is again independent of the systematic monetary policy parameters m_0 and γ, so that changes in these parameters will not change y_t. Hence, the policy invariance result also holds in this model. Note that this is a strong and seemingly surprising result: monetary policy cannot correct for the output gap that arises because of persistence in the model. Even if $y_t < y^{\mathrm{f}}{}_t$, an expansionary systematic monetary policy cannot reduce the output gap, nor increase it. Further, the sources of the output gap are persistence, random disturbances and fluctuations in the full-employment output, with actual output lagging behind the full-employment level. These provide the basis for the real business cycle (RBC) theories. Given that $0 < \alpha < 1$, the deviations of output from its full-employment level are self-correcting over time.

For another – though misleading – pattern of derivation in the above model, we have from (17):

$$\mathrm{E}m_t = m_0 + \gamma \mathrm{D}y_{t-1} \tag{21}$$

be at point c after the current period. Therefore, the increase in output is likely to be short lived. Its duration will depend upon the time it takes the public to correct for erroneous price expectations.

so that, from (16), (18) and (21):

$$\begin{aligned} \mathrm{D}y_t &= \alpha \mathrm{D}y_{t-1} + \beta m_t - \beta(m_0 + \gamma \mathrm{D}y_{t-1}) + \mu_t \\ &= (\alpha - \beta\gamma)\mathrm{D}y_{t-1} - \beta m_0 + \beta m_t + \mu_t \end{aligned} \tag{22}$$

In (22), $\mathrm{D}y_t$ depends upon the policy parameters m_0 and γ, so that we get the impression that output will depend upon systematic monetary change. However, this would be erroneous since substituting (17) in (22) to eliminate m_t again gives (20), which establishes the systematic policy irrelevance result.

17.3 The Lucas critique of estimated equations as a policy tool

Suppose the economic researcher were to use (20) to set up an estimating equation of the form:

$$y_t = a_0 + a_1 \Delta Y_t + \xi_t \qquad a_1 \geq 0 \tag{23}$$

where y is output, Y is nominal aggregate demand and ξ is white noise, and found the estimated value of a_1 to be positive (i.e. $\hat{a}_1 > 0$). As argued above, if the policy maker increased aggregate demand by more than was experienced during the estimation period, a_1 would shift, so that $\hat{a}_1$ would no longer be the relevant magnitude under the revised demand policy. Hence, a maintained shift in the expansion of demand does not leave the estimated parameters constant. This is known as the *Lucas critique* of estimated functions of the Lucas–Phillips curve type; if the underlying model of the economy is as set out by Lucas, the parameters of functions such as (23) are not invariant with respect to policy shifts.

Lucas (1973) estimated a variant of (23) for a cross section of countries and found that countries with low rates of inflation (such as the USA in the 1950s and 1960s) showed evidence of a positive relationship between output and demand increases while those with hyper-inflation (such as Argentina in the 1950s and 1960s) did not. He thereby concluded that this relationship shifts as inflation increases, and that the Phillips curve tradeoff does not hold at high and persistent rates of inflation.

Similarly, an estimating equation with unemployment u as the dependent variable would be:

$$u_t = b_0 + b_1 \Delta Y_t + \xi_t \qquad b_1 \leq 0 \tag{24}$$

Note that an estimated function such as (24) does not distinguish between the expectations-augmented Phillips curve, based on errors in price level expectations in labor markets, and the Lucas supply rule, based on errors in relative commodity prices, so that the estimated coefficients would reflect the influence of both types of errors. Equations such as (23) with output as the dependent variable or (24) with unemployment as the dependent variable have been estimated for a variety of countries and for a variety of time periods. Lucas's conclusion, that such functions are often unstable under demand policy shifts, is now well established.[6]

6 Another aspect of the preceding discussion and of the Lucas critique is that economic agents learn from experience, so that their expectations shift if policy shifts. This brings us back to the question of the expectations hypotheses and the role of learning in them. A number of learning mechanisms have been proposed and the speed at which expectational errors are eliminated is derived for them. However, they go beyond the concerns of this book and are not examined herein.

However, another of Lucas's conclusions was that systematic demand increases do not change real output and unemployment. This has not always been supported in empirical studies, as the stylized facts on money and output at the beginning of Chapter 14 show.

The Lucas critique in the LSW model

To check on the applicability of the Lucas critique in the LSW model comprising equations (16) to (18), restate (22) compactly as:

$$y_t = a_0 + a_1 y_{t-1} + a_2 m_t + \mu_t \tag{25}$$

where $a_0 = y^{\mathrm{f}}{}_t - (\alpha - \beta\gamma) y^{\mathrm{f}}{}_{t-1} - \beta m_0$, $a_1 = (\alpha - \beta\gamma)$ and $a_2 = \beta$. If (25) is estimated, it would yield the estimated values of a_1 and a_2 as $\hat{a}_1$ and $\hat{a}_2$. If $\hat{a}_2 > 0$, it would be tempting to conclude that the authorities could increase the money supply to increase output. However, as we have shown earlier in Section 17.2, this would not be a valid conclusion. A policy change in the money supply rule would mean a shift in the values of m_0 or/and γ. But, as these shift, a_1 and a_2 in (25) would shift, as can be seen from a glance at (22), of which (25) is only the compact form. Hence, it cannot be assumed that a_1 and a_2 are invariant to a policy change. The Lucas critique therefore applies to (25), so that this equation cannot be used as a tradeoff for policy formulation. Further, (25) cannot be used for regression estimates across policy regimes since such estimation assumes constant parameters.[7]

17.4 Testing the effectiveness of monetary policy: estimates based on the Lucas and Friedman supply models

The Friedman–Lucas supply function can be stated with output, employment, unemployment or another real variable, such as the real rate of interest as the dependent variable, and with the expectational errors in absolute prices, in aggregate demand or with just money supply as the independent variable. From a monetary policy perspective, the form of the Friedman–Lucas supply function that is usually tested is:

$$y_t = a_0 + a_1(M_t - M^{\mathrm{e}}{}_t) + \Sigma_j a'_j z_{jt} + \mu_t \qquad a_1 > 0 \tag{26}$$

where:

M = nominal money supply
M^{e} = expected nominal money supply
z_j = other exogenous variables
μ = random term.

Equation (26) focuses on money supply as the sole policy variable determining aggregate demand. Under rational expectations,

$$M^{\mathrm{e}}{}_t = \mathrm{E}M_t \tag{27}$$

7 For consistent and unbiased estimates using data from a given policy regime – i.e. with constant true values of m_0 and γ – estimate (21) and (25). Since there are five reduced form coefficients (a_0, a_1, a_2, m_0 and γ) in equations (24) and (26), while there are only four structural ones (m_0, γ, α, β) in (16) to (18), cross-equation restrictions implied by (25) will have to be imposed and a simultaneous estimation procedure used.

where EM_t is proxied by its estimated value $EM_t = \hat{M}_t$. Similarly, if needed, the estimated forms of the other variables in (26) can be used to modify this equation, but are omitted here. Therefore, the estimated form of (26) becomes:

$$y_t = a_0 + a_1(M_t - \hat{M}_t) + \Sigma_j a_j' z_{jt} + \mu_t \qquad a_1 > 0 \tag{28}$$

Equation (28) is a commonly used form of the modern classical output hypothesis for the short run. In the long run, $M = \hat{M}$, so that money supply changes cannot affect y.

17.4.1 A procedure for segmenting the money supply changes into their anticipated and unanticipated components

To estimate (28), we need the value of M^e_t. Its value is usually the one specified by the REH and is a function of the information available to the economic agent. Assuming that the central bank controls the national money supply, the relevant knowledge would be that of the public on central bank behavior and on the policy rule that the central bank follows. Assume that this is a rule that gives the money supply function as:

$$M_t = \Sigma_i \alpha_i x_{it} \tag{29}$$

where x_t is a set of exogenous and predetermined variables. Adding in a disturbance term η gives the money supply rule function:

$$M_t = \Sigma_i \alpha_i x_{it} + \eta_t \tag{30}$$

Under the REH, the public is assumed to know the policy function (29) and use the estimated values $\hat{\alpha}_i$ from (30) to calculate the estimated value $\hat{M}_t$, where $\hat{M}_t$ is the rational expectations' proxy for the anticipated money supply, so that:

$$\hat{\eta}_t = M_t - \hat{M}_t \tag{31}$$

$\hat{\eta}_t$ is the proxy, under the rational expectations hypothesis, for the unanticipated money supply.

The nested form of the Lucas model

The nested form of (28) and (26), incorporating both $\hat{\eta}_t$ and $\hat{M}$, specifies the linear (or log-linear) estimating equation as:

$$y_t = \beta_0 + \beta_1 \hat{M}_t + \beta_2 \hat{\eta}_t + \Sigma_j \gamma_j z_{jt} + \mu_t \tag{32}$$

If $\hat{\beta}_1 = 0$, the anticipated values of money supply do not affect real output, so that this finding

would be consistent with the modern classical hypothesis; but if $\hat{\beta}_1 > 0$, the modern classical hypothesis is rejected.[8,9]

Barro's test of the Lucas model: a joint test of neutrality and rational expectations

Barro (1977), in one of the earliest articles applying the REH to the Friedman–Lucas supply rule, used a two-step OLS procedure to test jointly for rational expectations and neutrality. In the first stage, he estimated by OLS a forecasting equation for the money supply. Under the assumption of rational expectations, the calculated value of the money supply from this estimation was used as the proxy for its anticipated value and the residual was used as the unanticipated value. The impact of these on the dependent real variable – in his case, unemployment – was then estimated in a second equation. Using this procedure, Barro (1977) reported the following estimated functions, using annual data for the USA for 1946–73:

$$\ln[U/(1-U)]_t = -3.07 - 5.8\text{DMR}_t - 12.1\text{DMR}_{t-1} - 4.2\text{DMR}_{t-2} - 4.7\text{MIL}_t + 0.95\text{MINW}_t \tag{33}$$

$$\text{D}\hat{M}_t = 0.087 + 0.24\text{DM}_{t-1} + 0.35\text{DM}_{t-2} + 0.082 \ln \text{FEDV}_t + 0.027 \ln [U/(1-U)]_{t-1} \tag{34}$$

where:

DM	= growth rate of money supply
D$\hat{M}$	= estimated growth rate of money supply
DMR	= unanticipated money growth rate (=DM $-$ D$\hat{M}_t$)
MIL	= military size
MINW	= minimum wage
FEDV	= federal government expenditures relative to their normal level
U	= unemployment rate.

Equation (33) is a version of the Friedman–Lucas supply rule and (34) is a money supply rule. Barro's estimates showed that unanticipated money growth was significant in explaining

8 $\beta_1 < 0$ indicates that increases in the money supply are detrimental for output in the economy, as when they increase the degree of uncertainty and reduce investment or lead otherwise to a diversion of resources to less efficient uses in the economy.

9 A simplified form of this system can give erroneous results, as Mishkin (1982) showed. Consider the following simple system:

$$y_t = a_1 \hat{M}_t + \Sigma_j g_j z_{jt} + \theta_t \tag{1}$$

$$M_t = b_1 y_{t-1} + \Psi_t \tag{2}$$

which imply that:

$$y_t = a_1 b_1 y_{t-1} + \Sigma_j g_j z_{jt} + \theta_t \tag{3}$$

where $\hat{M}_t$ does not occur as an explanatory variable in (3), so that it would appear that (3) rejects the Keynesian hypothesis when in fact this hypothesis was part of the initial model in the form of (1). This problematic result arose because (2) involved only the lagged terms of y as explanatory variables, so that (1) and (2) were not identified through the reduced form (3), thereby making it impossible to judge from (3) whether the anticipated part of the money supply affected real output or not. Hence, estimating (3) does not allow us to discriminate between the two hypotheses.

current unemployment. They also showed that when the total money supply, current and with two lags, replaced the unanticipated money supply terms in (33), their coefficients were not significant. Barro concluded that his study supported the modern classical hypothesis based on Lucas (1972, 1973) and not the Keynesian one.

17.4.2 Separating neutrality from rational expectations: Mishkin's test of the Lucas model

Mishkin (1982) objected to Barro's estimation procedure since it only provided a test of the joint hypotheses of the neutrality of money and rational expectations, without providing separate results on each of these hypotheses.[10] To understand Mishkin's objection, note that Barro's estimation system was of the form:

$$M_t = \Sigma_i \alpha_i x_{it} + \eta_t \tag{35}$$

$$y_t = \beta_0 + \sum_{j=0}^{n} \beta_j (M_{t-j} - \Sigma_i \alpha_i x_{it-j}) + \mu_t \tag{36}$$

where x_{it} were a set of exogenous or predetermined variables for determining the money supply. The output equation (36) embodies both neutrality and rational expectations. It also allows lags in the impact of the unanticipated money supply. Determinants of output other than the money supply have been left out of this equation, for simplification. The money supply equation (35) is essentially the same as (30), while (36) has been obtained by substituting the estimate of M_t from (30) in (28). This system imposes rational expectations since α_i in the money equation (35) also appears in the output equation (36). The neutrality property is imposed in (36) since the coefficients on EM_t are a priori set at zero.

To test for rational expectations and neutrality separately, the estimation system should be:

$$M_t = \Sigma_i \alpha_i x_{it} + \eta_t \tag{37}$$

$$y_t = \beta_0 + \sum_{j=0}^{n} \beta_j (M_{t-j} - \Sigma_i \alpha^*_i x_{it-j}) + \sum_{j=0}^{n} \gamma_j \alpha^*_{i,t-j} x_{i,t-j} + \mu_t \tag{38}$$

where (38) is the nested equation (32). Rational expectations require $\alpha^*_i = \alpha_i$, while neutrality requires $\gamma_j = 0$.

Therefore, maintaining the rational expectations hypothesis – that is, setting $\alpha_i = \alpha^*_i$ – while relaxing the neutrality assumption implies testing the system:

$$M_t = \Sigma_i \alpha^*_i x_{it} + \eta_t \tag{39}$$

$$y_t = \beta_0 + \sum_{j=0}^{n} \beta_j (M_{t-j} - \Sigma_i \alpha_i x_{it-j}) + \sum_{j=0}^{n} \gamma_j \alpha_{i,t-j} x_{i,t-j} + \mu_t \tag{40}$$

10 Mishkin also argued that while the two-step OLS estimation procedure will yield consistent parameter estimates, they do not generate valid F-test statistics, thereby resulting in inconsistent estimates of the standard errors of the parameters and test statistics, which do not follow the assumed F-distribution. He used the Full Information Maximum Likelihood (FIML) procedure for the nonlinear joint estimation for his systems.

The null hypothesis of neutrality – that is, $\gamma_j = 0$ – can be tested by comparing the estimates of the system (39) and (40) with those from (35) and (36). Maintaining the neutrality hypothesis while testing for rational expectations requires estimating:

$$M_t = \Sigma_i \alpha_i x_{it} + \eta_t \tag{41}$$

$$y_t = \beta_0 + \sum_{j=0}^{n} \beta_j (M_{t-j} - \Sigma_i \alpha^*_i x_{it-j}) + \mu_t \tag{42}$$

The null hypothesis of $\alpha^*_i = \alpha_i$ is tested by comparing the estimates from (41) and (42) against those from (35) and (36).

Mishkin's tests for the USA on quarterly data for 1954–76 used unemployment and output as the dependent variables, and nominal GNP and the rate of inflation among the independent variables. He reported that while the REH was not rejected by the data, neutrality was. Further, the estimated coefficients of the anticipated and unanticipated demand variables were very similar in magnitude. Therefore, Mishkin's results supported the Keynesian hypothesis and rejected the modern classical one on the key issue of the neutrality of anticipated aggregate demand and of demand management policies. These results did not reject the REH, which is merely a procedure for modeling expectations and, as noted earlier, is not a priori inconsistent with either the Keynesian or the modern classical theories.

Among several other studies, Frydman and Rappoport (1987) also tested for the distinction between the impact of anticipated and unanticipated monetary policy by examining the impact of the growth rate of money on output. Their findings reject this distinction for the short-run determination of output, with this rejection robust to different specifications of rational expectations and of the employment level of output.

We conclude that there is by now very substantial evidence that this distinction is not valid and, further, that anticipated monetary policy is not neutral, at least for the short run. These findings are not surprising in view of the stylized facts listed at the start of Chapter 14. Therefore, models of the Lucas and the Sargent and Wallace variety do not provide a useful basis for macroeconomic analysis and policy.

17.5 Distinguishing between the impact of positive and negative money supply shocks

It is sometimes argued that decreases in the money supply are likely to have a stronger impact than increases in it. There can be several reasons for this. Among these are:

1. A decrease in the money supply represents a decrease in credit in the economy, so that borrowers are forced to curtail their economic activities, and this reduces output in the economy. By comparison, an increase in the money supply means a greater willingness by the financial intermediaries to lend, which does not result in the same urgency to borrow as a decrease in loans to repay.[11]
2. Contractionary policies are likely to be pursued during booms with full employment and a high demand for investment and additional borrowing, whereas expansionary policies are likely to be pursued during recessions when firms generally face inadequate demand

11 An analogy sometimes used is that one can pull on a string but not push on it.

for their products, often have excess capacity and do not have enough incentive to increase their investment and borrowing. That is, the impact of the two types of policies also depends on the phase of the business cycle with which the two are associated.

3. The economy is likely to possess some downward rigidity in prices and nominal wages whereas it possesses a higher degree of flexibility for increases in them. Decreases in the money supply run into this downward rigidity and are more likely to have real effects, while most or all of the impact of increases in the money supply could be only on prices and the nominal but not the real value of output.

The Barro and Mishkin tests can be modified to test for this differential impact. We illustrate this by modifying the Friedman–Lucas supply function to:

$$y_t = a_0 + a^{+}{}_1 M^{\mathrm{u}}{}_t + \Sigma_j a'_j z_{jt} + \mu_t \tag{43}$$

and its nested form (32) to:

$$y_t = a_0 + a^{+}{}_1 M^{\mathrm{u}+}{}_t + a^{-}{}_1 M^{\mathrm{u}-}{}_t + \beta^{+}{}_1 M^{\mathrm{e}+}{}_t + \beta^{-}{}_1 M^{\mathrm{e}-}{}_t + \Sigma_j \alpha_j z_{jt} + \mu_t \tag{44}$$

where:

$M^{\mathrm{u}+}$ = unanticipated increase in the money supply
$M^{\mathrm{u}-}$ = unanticipated decrease in the money supply
$M^{\mathrm{e}+}$ = expected (anticipated) increase in the money supply
$M^{\mathrm{e}-}$ = expected (anticipated) decrease in the money supply

and z_j are the other variables in the determination of output. The other aspects of the estimation procedures of Barro and Mishkin remain as specified earlier. In particular, the actual estimation should allow for lags and other independent variables. The Mishkin (1982) procedure specified above can be suitably modified to check on neutrality, rational expectations and asymmetrical effects.

Some empirical studies do report evidence of the asymmetrical effects – and non-neutrality – of monetary policy. Our earlier arguments suggest the possibility that negative shocks to money have greater impact than positive ones. For example, Cover (1992), using Mishkin's (1982) procedure, finds for US quarterly data that positive shocks to M1 had no effect on output, whereas negative ones did so. Ratti and Chu (1997) confirm for Japan the asymmetry between the effects of positive and negative shocks. They further report that unanticipated changes in a wide definition of money did not have a significant impact on output in Japan.

17.6 LSW model with a Taylor rule for the interest rate

As Chapter 13 pointed out, in recent decades many central banks have chosen to set the interest rate rather than the money supply. Assume that the central bank uses a contemporaneous Taylor rule with price level targeting[12] of the form:

$$r^{\mathrm{T}}{}_t = r_0 + \lambda_y (y_t - y^{\mathrm{f}}) + \lambda_P (P_t - P^{\mathrm{T}}) \qquad \lambda_y, \lambda_P > 0 \tag{45}$$

There are two ways of integrating this rule in the LSW framework. One is to use the money demand function to derive the endogenous money supply that keeps the financial markets in

12 This has been done to avoid having the price level in some equations and the inflation rate in others within our model. P is the log of the price level and T indicates its target value.

equilibrium (see Chapter 13) and then use the resulting money supply function in the rest of the LSW model. The other method is to replace the money supply function in the original LSW model by the Taylor rule and make appropriate changes in the aggregate demand equation. The following derivations are based on the second method.

For the linear forms of the various functions, the IS equation for the open economy derived in Chapter 13 was:

$$y^{\mathrm{d}} = \alpha[\{c_0 - c_y t_0 + i_0 - i_r r + g + x_{c0} - x_{c\rho}\rho^{\mathrm{r}}\} + (1/\rho^{\mathrm{r}}).\{-z_{c0} + z_{cy}t_0 - z_{c\rho}\rho^{\mathrm{r}}\}] \quad (46)$$

where:

$$\alpha = \left(\frac{1}{1 - c_y + c_y t_y + \frac{1}{\rho^{\mathrm{r}}} z_{cy}(1 - t_y)}\right) > 0$$

and $\rho^{\mathrm{r}} = \rho P / P^{\mathrm{F}}$. The definitions of the symbols are as given in Chapter 13. For simplification, assume the general form of the IS equation to be:

$$y_t^{\mathrm{d}} = y_0 - \theta_1 r^{\mathrm{T}} - \theta_2 P_t \quad (47)$$

The resulting LSW model with the stochastic forms of the preceding IS equation and the Taylor rule and all variables in logs is:

$$\mathrm{D}y^{\mathrm{s}}{}_t = \alpha \mathrm{D}y_{t-1} + \beta(P_t - P^{\mathrm{e}}{}_t) + \mu_t \qquad \alpha, \beta > 0 \quad (48)$$

$$y^{\mathrm{d}}{}_t = \theta_0 - \theta_1 r_t - \theta_2 P_t + \nu_t \qquad \theta_1, \theta_2 > 0 \quad (49)$$

$$r_t = r^{\mathrm{T}}{}_t + \eta_{1t} \quad (50)$$

$$r^{\mathrm{T}}{}_t = r_0 + \lambda_y \mathrm{D}y_t + \lambda_P(P_t - P^{\mathrm{T}}) + \eta_{2t} \qquad \lambda_y, \lambda_P > 0 \quad (51)$$

$$y_t = y^{\mathrm{d}}{}_t = y^{\mathrm{s}}{}_t \quad (52)$$

$$P^{\mathrm{e}}{}_t = \mathrm{E}P_t \quad (53)$$

where $\mathrm{D}y = y_t - y^{\mathrm{f}}$, μ is supply shocks, ν is IS shocks, η_2 is the monetary policy shock and η_1 is the stochastic slippage in the control of the economy's interest rate by the central bank. All disturbances are taken to be white noise.

Taking the rational expectation of the aggregate supply equation (48), we have:

$$\begin{aligned} \mathrm{ED}y^{\mathrm{s}}{}_t &= \alpha \mathrm{ED}y_{t-1} + \beta[\mathrm{E}P_t - \mathrm{E}(\mathrm{E}P_t)] + \mathrm{E}\mu_t \\ &= \alpha \mathrm{D}y_{t-1} \end{aligned} \quad (54)$$

so that:

$$\mathrm{D}y^{\mathrm{s}}{}_t - \mathrm{ED}y^{\mathrm{s}}{}_t = \beta(P_t - \mathrm{E}P_t) + \mu_t \quad (55)$$

Further, from (48) and (53):

$$P_t - \mathrm{E}P_t = (1/\beta)(\mathrm{D}y_t - \alpha \mathrm{D}y_{t-1}) - (1/\beta)\mu_t \tag{56}$$

Now, taking the rational expectation of the IS equation (49), we have:

$$\mathrm{E}y^{\mathrm{d}}{}_t = \theta_0 - \theta_1 \mathrm{E}r_t - \theta_2 \mathrm{E}P_t \tag{57}$$

Subtracting (57) from (49) yields:

$$y^{\mathrm{d}}{}_t - \mathrm{E}y^{\mathrm{d}}{}_t = -\theta_1(r_t - \mathrm{E}r_t) - \theta_2(P_t - \mathrm{E}P_t) + \nu_t \tag{58}$$

where r_t is given by:

$$r_t = r_0 + \lambda_y \mathrm{D}y_t + \lambda_P(P_t - P^{\mathrm{T}}) + \eta_{2t} + \eta_{1t} \tag{59}$$

so that:

$$\mathrm{E}r_t = r_0 + \lambda_y \mathrm{E}\mathrm{D}y_t + \lambda_P(\mathrm{E}P_t - P^{\mathrm{T}}) \tag{60}$$

Hence:

$$r_t - \mathrm{E}r_t = \lambda_y(\mathrm{D}y_t - \mathrm{E}\mathrm{D}y_t) + \lambda_P(P_t - \mathrm{E}p_t) + \eta_{2t} + \eta_{1t} \tag{61}$$

Substituting (61) in (58) and imposing equilibrium, so that $y_t = y^{\mathrm{d}}{}_t = y^{\mathrm{s}}{}_t$, and $\mathrm{D}y_t = \mathrm{D}y^{\mathrm{d}}{}_t = \mathrm{D}y^{\mathrm{s}}{}_t$, we have:

$$y_t - \mathrm{E}y_t = -\theta_1\{\lambda_y(\mathrm{D}y_t - \mathrm{E}\mathrm{D}y_t) + \lambda_P(P_t - \mathrm{E}P_t) + \eta_t\} - \theta_2(P_t - \mathrm{E}P_t) + \nu_t \tag{62}$$

where $\eta_t = \eta_{1t} + \eta_{2t}$. Since $y_t - \mathrm{E}y_t = \mathrm{D}y_t - \mathrm{E}\mathrm{D}y_t$, where $\mathrm{D}y = y - y^{\mathrm{f}}$, (62) becomes:

$$\mathrm{D}y_t = \mathrm{E}\mathrm{D}y_t - \theta_1\lambda_y \mathrm{D}y_t + \theta_1\lambda_y \mathrm{E}\mathrm{D}y_t - \theta_1\lambda_P(P_t - \mathrm{E}P_t) - \theta_1\eta_t - \theta_2(P_t - \mathrm{E}P_t) + \nu_t \tag{63}$$

$$\begin{aligned}(1 + \theta_1\lambda_y)\mathrm{D}y_t &= (1 + \theta_1\lambda_y)\mathrm{E}\mathrm{D}y_t - (\theta_1\lambda_P + \theta_2)(P_t - \mathrm{E}P_t) - \theta_1\eta_t + \nu_t \\ &= [\{(1 + \theta_1\lambda_y)\alpha \mathrm{D}y_{t-1} - (1/\beta)(\theta_1\lambda_P + \theta_2)\{(\mathrm{D}y_t - \alpha \mathrm{D}y_{t-1}) - \mu_t\} - \theta_1\eta_t + \nu_t \\ \{(1 + \theta_1\lambda_y) + (1/\beta)(\theta_1\lambda_P + \theta_2)\}\mathrm{D}y_t &= \{(1 + \theta_1\lambda_y) + (1/\beta)(\theta_1\lambda_P + \theta_2)\}\alpha \mathrm{D}y_{t-1} - \{(1/\beta)(\theta_1\lambda_P + \theta_2)\}\mu_t - \theta_1\eta_t + \nu_t\end{aligned} \tag{64}$$

Replacing $\{(1 + \theta_1\lambda_y) + (1/\beta)(\theta_1\lambda_P + \theta_2)\}$ by a_1 and $\{(1/\beta)(\theta_1\lambda_P + \theta_2)\}$ by a_2, we get:

$$\mathrm{D}y_t = \alpha \mathrm{D}y_{t-1} - (a_2/a_1)\mu_t - (1/a_1)\theta_1\eta_t + (1/a_1)\nu_t \tag{65}$$

where α is the fraction by which the economy, on its own, adjusts the current period's output gap as a fraction of last period's gap. It is independent of the systematic policy parameters λ_y, λ_P embodied in the Taylor rule. However, random IS shocks (μ), monetary sector shocks (η_1) and policy shocks (η_2) do impact on the output gap, with this impact a function of various parameters, including those of policy. This conclusion is similar to that derived earlier from

the LSW model with a money supply function. In particular, systematic policy parameters do not affect output in both models. This is especially surprising in the current model with the Taylor rule, which explicitly targets the output gap. This finding highlights the point that, in this model, the driving conclusions on output are given by the nature of the supply rule and the expectations hypothesis, not by the monetary policy rule, which determines systematic monetary policy. In the LSW model, this supply rule is the Lucas rule. Replacing it by a Keynesian or NK supply function is essential for showing the effectiveness of systematic policies in changing the output gap.

Note that the policy parameters do affect the impact on output of each of the disturbances.

17.7 Testing the effectiveness of monetary policy: estimates from Keynesian models

17.7.1 Using the LSW model with a Keynesian supply equation

A Keynesian supply function would differ from the Lucas one in several ways. One, it would allow both anticipated and unanticipated money supply changes to affect output. Two, as argued in Chapter 15, the effect of money supply changes on output would depend on the state of the economy and the degree of involuntary unemployment. Three, in Keynesian models, the impact of money supply changes on output can occur without a prior change in the price level, so that the LSW(II) model with the money supply as an explanatory variable is preferable to the LSW model with the price level. Further, note that, in the Keynesian context, the dependence of y_t on y_{t-1} occurs because Keynesian models allow for staggered wage contracts longer than one period as well as the gradual adjustment of output to shocks.

Adapting the Friedman–Lucas supply function to a Keynesian format by replacing the unanticipated money supply ($m_t - m^e_t$) by the total money supply m_t gives:

$$\mathrm{D}y_t = \alpha y_{t-1} + \beta_t m_t + \mu_t \qquad \alpha, \beta > 0 \tag{66}$$[13]

where the definitions of the symbols are as given earlier. As a reminder, note that the lower-case letters are logs of variables and Dy is the deviation from the full-employment level and not a change from the previous period's level. For Keynesian models, since the money multiplier depends on the state of the economy, the value of β_t should be a function of the output gap.

If we specify the complete model as consisting of the money supply function (17) and the Keynesian output supply function (66), we find that:

$$\mathrm{D}y_t = (\alpha + \beta_t\gamma)\mathrm{D}y_{t-1} + \beta_t m_0 + \mu_t + \beta_t\xi_t \tag{67}$$

where Dy_t depends upon the policy parameters m_0 and γ, which can be used to achieve the desired objectives with respect to y_t, and upon the money multiplier β_t, as well as on the previous period's deviation from full-employment output.

Equation (67) is not an appealing format for the Keynesian supply function. A better format seems to be one where the deviation of output from its full-employment level depends not

13 β in true Keynesian models would properly not be a constant but would be a function of the deviation of output from full-employment level. However, Sargent and Wallace (1976) proposed $\mathrm{D}y_t = \alpha\mathrm{D}y_{t-1} + \beta m_t + \mu_t$ as the Keynesian reduced-form equation, with a constant β, and it is frequently cited as such.

on the money supply but on its change. The output supply function consistent with this idea is:

$$Dy_t = \alpha Dy_{t-1} + \beta_t(m_t - m_{t-1}) + \mu_t, \qquad \alpha, \beta > 0 \tag{68}$$

The LSW money supply rule (17) can also be reformulated as:

$$m_t - m_{t-1} = \gamma_1(y_{t-1} - y_{t-2}) - \gamma_2(y^{f}{}_{t-1} - y^{f}{}_{t-2}) + (\xi_t - \xi_{t-1}) \qquad \gamma_1, \gamma_2 < 0 \tag{69}$$

This modification of the LSW money supply rule allows for differential response of the money supply to changes in actual output versus changes in full-employment output. Since $\gamma_1, \gamma_2 < 0$, the monetary authority decreases the money supply if the actual output rose last period and increases it if there was a rise in the full-employment output.

Equations (68) and (69) yield:

$$Dy_t = \alpha Dy_{t-1} + \beta_t\gamma_1(y_{t-1} - y_{t-2}) - \beta_t\gamma_2(y^{f}{}_{t-1} - y^{f}{}_{t-2}) + \mu_t + \beta_t(\xi_t - \xi_{t-1}) \tag{70}$$

which implies that the monetary authority can change its policy parameters γ_1, γ_2 to affect output deviations from full employment. Therefore, policy irrelevance does not occur in this Keynesian model, irrespective of any assumption on the rationality of expectations. However, the Lucas critique will still apply since the coefficient of y_{t-1} in (70) depends on the monetary policy parameters γ_1 and γ_2, which will change with a policy shift.

17.7.2 Gali's version of the Keynesian model with an exogenous money supply

Chapter 15 reported the findings from Mishkin's (1982) test from reduced-form equations that money was not neutral while rational expectations were valid.

Gali (1992) uses a structural model that is Keynesian with a Phillips curve. His model, in logs, is:

IS equation:

$$y_t = \mu^{s}{}_t + \alpha - \sigma(R_t - E\Delta p_{t-1}) + \mu^{IS}{}_t \tag{71}$$

LM equation:

$$m_t - p_t = \phi y_t - \lambda R_t + \mu^{md}{}_t \tag{72}$$

Money supply process:

$$\Delta m_t = \mu^{ms}{}_t \tag{73}$$

Phillips curve:

$$\Delta p_t = \Delta p_{t-1} + \beta(y_t - \mu^{s}{}_t) \tag{74}$$

where the symbols designate the logs of the relevant variables, except for R (which stands for the level of the nominal interest rate). μ^{s}, μ^{IS}, μ^{md} and μ^{ms} are the stochastic processes for output supply, expenditures, money demand and money supply respectively. The IS and LM equations are consistent with the Lucas model, except that they provide greater detail.

The major difference between this Keynesian model and the LSW models lies in the specification of output supply. Gali specifies it by the Phillips curve, so that changes in the inflation rate between periods determine the variations μ^s in output from its equilibrium level, whereas the LSW model uses relative price misperceptions to explain such deviations. In the Gali model, output can change due to supply shocks through μ^s, or demand shocks due to μ^{IS}, μ^{md} or μ^{ms}. Positive demand shocks increase both output and prices while positive supply shocks increase output and decrease prices. Monetary shocks are transmitted to the real sector only through changes in the interest rate.

The segregation of the experienced shocks into four different types requires special assumptions on their origin or impact. Gali separated the supply shocks from the demand shocks by the assumption that the former have long-run effects on output while the latter do not. The IS shocks were separated from the money market shocks by the assumption that the latter do not have contemporaneous impact on aggregate demand in the same quarter, since their impact occurs through the changes in interest rates impinging on investment. The money demand and supply shocks were separated under three alternative assumptions, which we do not report here.

Gali's data was quarterly for the USA for 1955:I to 1987:IV. The monetary aggregate used was M1 and the interest rate was represented by the three-month Treasury bill rate. The findings supported the Keynesian claim that demand shocks do cause output changes while rejecting its claim that they, rather than supply shocks, were the dominant source of output fluctuations. Supply shocks had a substantial deflationary impact and accounted for about 70 percent of output variability over the business cycle. However, their impact on the nominal interest rate was small.

Increases in M1 increased output, reaching a peak in about four quarters, accompanied by increases in inflation and nominal interest rates but decreases in the real rate. While the money supply shocks accounted for most of the short-run variability of the nominal and real interest rates and for some variability in output over the business cycle, there was no long-run effect on output or the real rate, though this result really emanates from the built-in assumptions, but only on the inflation rate. Money demand shifts had a much faster impact on prices than money supply shifts and significant impact on real balances, but little influence on output variability or the real rate.

The impact of the IS shocks on output started within the same quarter as the shock, clearly just a result of the assumptions made, reached a peak two quarters after the shock but almost vanished after four quarters. However, such shocks had permanent effects on money growth, inflation and the nominal rate. While they increased the nominal rate, they first increased the real rate but, because of their impact on inflation, soon led to a decline in the real rate, which returned to its initial level about two years after the shock. IS shocks accounted for a substantial part of the business fluctuations.

Gali's findings, therefore, support the Keynesian conclusions that demand shocks do cause variations in output. Further, money supply variations had a longer-lasting impact on output than IS shocks. However, the major source of the variations in output was supply rather than demand shocks. Demand shocks did not produce long-run effects on output and the real rate of interest.[14] While Gali's model did not distinguish between positive and negative

14 Note that these findings are affected by the assumptions made to segment the shocks to the economy, so that any errors in these assumptions could lead to erroneous results. For example, if aggregate demand changes do cause long-run changes in output, such impact would have been erroneously attributed to supply factors.

money supply shocks or test for asymmetry in their effects, it can clearly be modified to do so.

In another study, for a large number of countries, Bullard and Keating (1995) used vector autoregression to investigate the impact of inflation on output. In their procedure, they identified the permanent component of inflation as being due to permanent changes in the money growth rate while exogenous shocks to output were taken to cause only transitory shocks to inflation. Their finding was that permanent shocks to inflation did not *permanently* increase the level of output for most of the countries but did do so for certain low-inflation countries. In general, the estimated effects were positive for low-inflation countries, and low or negative for high inflation countries. Money is not neutral under these findings. This pattern of the effects of inflation on output seems to reflect the opinion of most monetary economists.

17.8 A compact form of the closed-economy new Keynesian model

The following presents a compact model of the new Keynesian model along the lines discussed in Chapter 15 (Bernanke and Woodford, 1997; Clarida *et al.* (Clarida, Gali and Gertler, CGG), 1999; Levin *et al.* 1999, 2001). It has three core equations: the IS equation, the output supply equation and the monetary policy rule.

IS equation:

$$x_t = E_t(x_{t+1}) - \psi(R_t - E\pi_{t+1}) + g_t \tag{75}$$

where $x = y - y^{\mathrm{f}}$, r is real interest rate while R is the nominal rate, π is the inflation rate and g represents all sources of expenditure (e.g. government deficits) other than investment. Optimizing forward-looking consumers and firms, the Fisher equation for perfect capital markets (so that $r_t = R_t - E\pi_{t+1}$) and rational expectations have been incorporated in this derivation of the IS equation.

Price adjustment process ("new Keynesian Phillips curve"):

$$\pi_t = \alpha x_t + \beta E_t \pi_{t+1} + v_t \tag{76}$$

where x is the output gap, acting as a proxy for marginal cost, mc, which rises with an increase in output. Firms rationally anticipate future inflation and smooth price adjustments. $v_t = \rho v_{t-1} + \eta_t$ and η is a random variable with a zero mean and constant variance. Besides other sources of disturbances, it can encompass deviations from the linear impact of the output gap on marginal cost.[15]

The preceding price adjustment equation can be rewritten as the output supply equation:

$$x_t = (1/\alpha)\pi_t - (\beta/\alpha)E_t\pi_{t+1} - (1/\alpha)v_t \tag{77}$$

In this format, the output gap responds to both current and future inflation.

15 See Gali and Gertler (1999), Clarida *et al*. (1999, p. 1667, fn 15).

The central bank's monetary policy rule:

The central bank derives its optimal interest rate rule by minimizing a quadratic loss function over inflation and the output gap. This yields the central bank's real interest rate rule as:

$$r^{\mathrm{T}}{}_t = r^{\mathrm{LR}} + \lambda x_t + \beta(E_t\pi_{t+1} - \pi^{\mathrm{T}}) \qquad \lambda, \beta > 0 \tag{78}$$

where r^{T} is the target interest rate and r^{LR} is the long-run interest rate.

The model consisting of equations (75), (77) and (78) clearly does not have short-run neutrality of monetary policy; a change in the interest rate by the central bank changes aggregate demand in the IS equation, which changes the inflation rate, which, in turn, changes the output gap. Since firms simultaneously determine their output and prices in response to changes in demand, an alternative interpretation of the sequence of effects would be: a change in the interest rate by the central bank changes aggregate demand in the IS equation, to which the response by firms changes their output and marginal costs, which leads to changes in their prices, so that the output gap and inflation change.

Long-run output and inflation:

In the long run, output is at its full-employment level which, by definition, is the long-run equilibrium level of output. Hence, by the assumptions for the long run, the output gap x^{LR} is zero, so that the long-run price adjustment equation becomes:

$$\pi^{\mathrm{LR}}{}_t = \beta E_t \pi^{\mathrm{LR}}{}_{t+1} + v_t \tag{79}$$

Note that with long-run output always equal to its full-employment level, which is independent of aggregate demand and its determinants, output is invariant with respect to monetary policy. Therefore, monetary policy can only affect the inflation rate. It affects the current inflation rate through the expected future rate, which depends on monetary policy and its credibility. In particular, a credible monetary policy with a target inflation rate of π^{T} will ensure that future inflation will be at this rate.[16]

17.8.1 Empirical findings on the new Keynesian model

Chapter 15 has already discussed the empirical validity of the new Keynesian model in general, and that of its three components. The following adds to the evidence discussed there and focuses on the components individually.

Empirical findings on the Taylor rule

The empirical validity of the Taylor interest rate rule depends on the primary monetary policy instrument used by the central bank. Chapter 13 had argued that it cannot a priori be taken for granted that the central bank sets the money supply or the interest. Some central banks set one and some the other one. For central banks that set the interest rate, there is by now substantial evidence that they follow some form, though often with time-variant coefficients, of the

16 Stationarity of the inflation rate in the long run requires that current inflation and future inflation be equal on average, which would require that $\beta = 1$.

Taylor rule, even when they do not explicitly announce this as their practice (see Chapter 15, Section 15.6). In general, the likelihood of a finding of the Taylor rule is very high for central banks that attempt to stabilize aggregate demand and control inflation through interest rates. However, since the coefficients of the Taylor rule depend on central bank preferences and the economy's constraints, there will be differences in the estimated coefficients of the Taylor rule among countries and even among different monetary policy regimes (e.g. with a different head of the central bank) of any given country.

In any case, central banks do not explicitly disclose their form of the Taylor rule and its coefficients for the output gap and the deviation of inflation from its target level, so that it needs to be estimated from *ex post* data. While some form of the Taylor rule does quite well empirically, its consistency with the new Keynesian models requires a forward-looking version. Levin *et al*. (1999, 2001) present the estimated coefficients of the Taylor rule from several studies, with the coefficients, and sometimes the rule itself, differing quite significantly between the studies. They conclude that a simple version of the inflation and output-targeting rule for the US economy performs quite well for the USA, and that a robust policy rule includes responses to a short-horizon forecast of inflation, not exceeding one year, and the current output gap. It also incorporates a high degree of policy inertia.

Maria-Dolores and Vazquez (2006) compare the performance of four (contemporaneous, backward-looking, a priori forward-looking, and the forward-looking rule derived from the central bank's optimization of its loss function) types of Taylor rule. They report that the new Keynesian model does much worse with a rule derived from the central bank's optimization than the backward-looking and the simple a priori forward-looking rules. Further, a simple autoregressive model of the interest rate can sometimes do better than the Taylor rule, as Depalo (2006) reports for Japan.

Empirical findings on the new Keynesian price adjustment equation

Rudd and Whelan (2003) test the general form of the forward-looking NK output equation:

$$\pi_t = \lambda x_t + \beta E_t \pi_{t+1} + v_t \tag{80}$$

where x can be specified as the output gap or the deviation of the unemployment rate from the natural one. They report that this equation performs very poorly for USA data. Empirically, current inflation is negatively, not positively, related to the future output gap; intuitively put, inflation is a negative leading indicator of future output. One reason for such a result could be that their proxy used for full-employment output is a poor one. A second reason could be that, since marginal cost is unobservable, the output gap in the NKPC was used as a proxy for real marginal cost (i.e. the ratio of marginal cost to price), but may be a poor proxy. Another proxy that has been suggested for the real marginal cost is labor's share in income (Gali and Gertler, 1999), but this also has not given much better results.

A third reason for the poor performance of the NKPC could be the high degree of persistence in inflation, which depends heavily on its own lagged values (Rudd and Whelan, 2003; Maria-Dolores and Vazquez, 2006). To capture this persistence, one suggestion is to replace the NKPC by a hybrid rule, say of the form:

$$\pi_t = \delta_1 \sum_{i=1}^{n} \pi_{t-i} + \alpha \sum_{j=0}^{\infty} \beta^j E_t x_{t+j} \tag{81}$$

A simpler form of this equation would use, for the lagged terms, only last period's inflation rate. However, for the NKPC, the justification for the backward-looking, rather than the forward-looking, inflation term on the right-hand side is problematical: for one thing, interpreting (81) as a form of (80) would be a denial of rational expectations. Further, since persistence plays a big part in inflation, the estimate α may not prove to be significant, so that (81) would reduce to just a form of static expectations. However, even if α proves to be significant, the estimated contribution of the forward output gaps may prove to be relatively small. Moreover, the estimated coefficients of this equation may shift with policy, so that the Lucas critique applies to them.

Further, as mentioned in Chapter 15, Mankiw (2001) provides three inconsistencies between this sticky price adjustment equation and the stylized facts about the relationship between inflation and unemployment. Among these inconsistencies are that this process does not generate persistence in inflation but does generate implausible dynamic adjustments in inflation and unemployment in response to monetary shocks. This study reports that the relationship that best fits the facts is the backward-looking one:

$$\pi_t = \beta\pi_{t-1} + \alpha(u_t - u^{\mathrm{n}}) + v_t \tag{82}$$

which is closer to the traditional Phillips curve, or to one with static or adaptive expectations. To provide a theoretical justification consistent with the current intertemporal optimizing approach in new Keynesian economics, Mankiw and Reis (2002, 2006a,b) replace the sticky price process by a sticky information one. The latter implies a backward-looking inflationary process which generates persistence in inflation, so that there is a significant difference between the sticky price adjustment and the sticky information equations of the Phillips curve.

17.8.2 Ball's Keynesian small open-economy model with a Taylor rule

Ball (1999, 2000) presents the following compact Keynesian model, with all variables in logs, for a small open economy:

IS equation:

$$y_t = -\beta r_{t-1} + \delta e_{t-1} + \lambda y_{t-1} + \varepsilon_t \tag{83}$$

Phillips curve:

$$\pi_t = \pi_{t-1} + \alpha y_{t-1} + \gamma(e_{t-1} - e_{t-2}) + \eta_t \tag{84}$$

Exchange rate determination:

$$e_t = \theta r_t + \upsilon_t \tag{85}$$

In these equations, y is the log of output, r is the real interest rate, e is the log of the real exchange rate[17] (a higher value of e means appreciation of the domestic currency), π is the current inflation rate and ε, η and υ are white noise terms. All parameters are taken to

17 For small open economies, the literature indicates that the central bank's policy function should include the exchange rate in addition to inflation and output among its variables (see, for example, Ball, 1999, 2000), though rules excluding the exchange rate also seem to do quite well.

be positive. (83) is an open-economy IS equation, with commodity demand depending on the (lagged values of the) real interest rate and the real exchange rate. (84) is an open-economy backward-looking price/inflation adjustment relationship, with the change in the inflation rate a function of lagged output and the lagged change in the exchange rate. The change in the exchange rate affects inflation through imports since depreciation increases import prices, which increases the domestic price level. (85) has a negative relationship between the interest rate and the exchange rate; an increase in the domestic interest rate increases capital inflows into the domestic country, which causes an appreciation of the domestic currency (i.e. the exchange rate rises). This model clearly embodies the short-run non-neutrality of aggregate demand and therefore of monetary policy, and is consistent with the new Keynesian sticky price hypothesis.

The preceding model is missing an equation for the money market. Such an equation can be the LM equation under the assumption of an exogenous money supply or an interest rate policy rule, such as the Taylor rule. Ball assumed that the central bank sets the real interest rate as:

$$r_t = a_0 r_{t-1} + a_\pi [\mathrm{E}(\pi_{t+4}|I_t) - \pi_t^*] + a_y y\mathrm{gap}_t + \mu_t \tag{86}$$

where $\pi_t^* = \pi_t - \gamma e_{t-1}$; that is, the central bank makes the desired long-run inflation rate invariant to changes in the exchange rate by filtering out the impact of lagged exchange rate changes on the current inflation rate.[18] y gap is the output gap, defined as the deviation of output from its full-employment level.

17.9 Results of other testing procedures

The findings in numerous empirical studies have ranged back and forth against the neutrality assumption. We do not intend to review many more studies but do consider the findings of a different type of study to be worth mentioning. As against the use of compact (small) reduced-form models reported above from the works of Barro, Mishkin, Gali, and Bullard and Keating, Mosser (1992) based her findings on four large structural macroeconometric models, well established in the late 1980s and 1990s.[19] These were used to generate the elasticities of real output, real interest rates and various other real variables with respect to monetary variables such as M1 and non-borrowed reserves held by banks. The estimated elasticities were significant not only for the first four quarters but for periods longer than 12 quarters for many of the real variables. For real output, the elasticities were positive and continued to increase up to 12 quarters, and were significant (either positive or negative) even at 40 quarters. Hence, not only were the monetary variables not neutral, the estimated lags were very long.

17.10 Summing up the empirical evidence on monetary neutrality and rational expectations

The results reported from Mishkin (1982), Gali (1992), Mosser (1992) and Ball (1999, 2000), as against Barro's (1977), supported the Keynesian theory and rejected the modern classical

18 To illustrate, for China, Wang and Handa (2007) estimate (83) to (85) by a simultaneous equations technique and estimate the Taylor-rule equation (86) using cointegration and error-correction modeling. They find support for both this rule and the above model for China, a developing economy.

19 These were: the Bureau of Economic Analysis model, the Data Resources Inc. model, the Federal Reserve Board/MPS model and the Wharton Econometric Forecasting Associates model.

one on the critical difference between them on the neutrality of anticipated money supply changes and the continual clearance of all markets. This implies rejection of the Friedman–Lucas supply rule, which incorporates the neutrality of anticipated money supply changes. In general, in spite of its optimizing microeconomic foundations and their intellectual appeal, the modern classical macroeconomic models have not been an unqualified empirical success.

Another aspect of neutrality modeling worth noting is that the supply functions used embody either neutrality (for classical models) or non-neutrality (for Keynesian models) of systematic monetary policy. However, empirical evidence suggests that systematic monetary policy can be neutral sometimes and non-neutral at other times, and that the "degree of non-neutrality" can be variant in intermediate cases, though it is a priori difficult to separate these cases (Lucas, 1994, 1996). These results need not come as a big surprise; the various theories presented in Chapters 14 and 15 indicate that there can be very many different causes of non-neutrality in the economy: staggered wage contracts and the lagged adjustment of nominal wages to prices; disequilibrium factors in the neoclassical model; deficient demand in the Keynesian model; relative price errors in a Lucas-type model; sticky prices in certain markets, sticky information and sticky pensions and other predetermined sources of incomes, etc. Consequently, if our estimating equation allows only the black/white scenario of neutrality versus non-neutrality, when part of the data is from a neutral sample and part is from a non-neutral one, we are likely to get a mixed bag of empirical results, varying with the relative weights of the two types of data in our sample.

Overall, while there is a somewhat mixed bag of empirical studies favoring one or the other of these hypotheses, the empirical evidence seems more often to favor the short-run non-neutrality of money rather than its neutrality. The evidence also seems to favor the finding that there is a difference between the effects of rationally anticipated *money supply* changes and those of unanticipated *price* changes, but both can have short-run effects on real output.

By comparison, the empirical evidence on the rationality of expectations does not, in general, reject it. In any case, its assertion that economic agents take account of all available information rather than merely that on the past, seems to be incontrovertible. However, at the level of implementation, whether this hypothesis justifies the interpretation of the modern classical approach that the expected values for the next few quarters or even years are the long-run equilibrium ones is highly doubtful. The more realistic interpretation clearly seems to be the Keynesian one: the rationally expected values are related to the actual future values, which depend on the actual performance of the economy, which is not necessarily the full-employment one, and the stage of the business cycle.

17.11 Getting away from dogma

The LSW model makes the implicit assumption, derived from perfectly competitive and efficient markets, that changes in output occur only in response to changes in prices. As some of the models argue, competitive markets do not necessarily have instantaneous restoration of prices to their equilibrium levels after a shock. If markets are slow to adjust but economic agents react faster to changes in demand or supply shocks, economic agents may change their output, employment, consumption and investment without a prior (or "full") change in prices and wages. Given the sluggish adjustment of prices and wages by markets, appropriate models for the economy must also consider the possibility of firms' responses to actual and expected changes in demand and workers' responses to the expected and actual changes in employment. In such a context, output and employment may respond to policy measures

without the impact of these policies first occurring through a price change. Chapter 14 supports this proposition by a quote from Robert Lucas, Jr, that is worth repeating here:

> Sometimes, as in the U.S. Great Depression, reductions in money growth seem to have large effects on production and employment. Other times, as in the ends of the post-World War I European hyperinflations, large reductions in money growth seem to have been neutral, or nearly so.
>
> (Lucas, 1994, p. 153).

> Anticipated and unanticipated changes in money growth have very different effects. [However, on the models that attribute this non-neutrality to unanticipated or random changes in the price level, the evidence shows that] *only small fractions* of output variability can be accounted for by unexpected price movements. Though the evidence seems to show that monetary surprises have real effects, *they do not seem to be transmitted through price increases*, as in Lucas (1972).
>
> (Lucas, 1996, p. 679, italics added).

These quotes, and the inconsistency of the implications of the Friedman–Lucas supply equation with the stylized facts set out at the beginning of Chapter 14, are convincing evidence that this supply equation is not valid for most or all modern economies.

17.11.1 The output equation revisited

Our preceding arguments and empirical assessment can be captured by using the identity:

$$y \equiv y^{\mathrm{f}} + (y^* - y^{\mathrm{f}}) + (y - y^*) \tag{87}$$

where y is the actual output, y^{f} is the long-run output in the absence of errors in expectations, and y^* is the short-run equilibrium unemployment rate in the presence of errors in expectations. The Friedman and Lucas supply analyses (see Chapter 14) imply that:

$$y^* - y^{\mathrm{f}} = f(P - P^{\mathrm{e}}) \tag{88}$$

Therefore,

$$y = y^{\mathrm{f}} + f(P - P^{\mathrm{e}}) + [(y - y^*)|\Delta Y/\Delta P] \tag{89}$$

where $[(y - y^*)|\Delta Y/\Delta P]$ is meant to indicate the deviation of actual output from the short-run value, resulting from errors in price expectations in the labor and commodity markets. Hence there are two reasons for deviations from full-employment output. One of these occurs through price changes. For these deviations, the Friedman and Lucas supply analyses imply that, for perfect markets and rational agents, it is not the price change itself that is relevant but the deviation of the price level from its expected value. The second reason for deviations arises from the changes in aggregate demand or supply that do not proceed through price level changes. These are captured through the term listed as $[(y - y^*)|\Delta Y/\Delta P]$, which has numerous potential causes and may occur in some stages of the economy but not in others. If none of them operate in a particular context, this term would be zero, so that anticipated changes in demand induced by anticipated monetary and fiscal policies would not have any

impact on output. But, as the Keynesians argue, the term need not be zero in all potential cases. Hence, expansionary monetary and fiscal policies may change the price level but their impact on output is unlikely to be fully reflected through price changes. In Lucas's assessment, only a fraction is so reflected, so that the larger part of the impact of money supply changes on output seems to occur through the term $[(y - y^*)|\Delta Y/\Delta P]$, whose determinants need to be specified. The Keynesian paradigm indicates several of them, but there may still be others.

While Keynesian models allow the relationship between output and inflation in period t to depend on expected inflation in future periods ($t+1$, and so on) they do not incorporate the determinants and effects of errors in expectations ($P_t - P^e_t$) or ($\pi_t - \pi^e_t$) during the current period. Classical models do so but suffer from their failure to incorporate the determinants and effects of $[(y - y^*)|\Delta Y/\Delta P]$ in their models. Addressing these deficiencies in a single model may help in addressing Lucas's criticism, quoted above, of existing macroeconomic models: there is no generally accepted model in which reductions in money growth sometimes "seem to have large effects on production and employment" while, at other times, "large reductions in money growth seem to have been neutral, or nearly so" (Lucas, 1994, pp. 153–4).

17.11.2 The Phillips curve revisited

For explaining the observed levels of unemployment, the preceding arguments imply that the appropriate form of the Phillips curve should be derived from the identity:

$$u \equiv u^n + (u^* - u^n) + (u - u^*) \tag{90}$$

where u^n is the natural rate of unemployment and u* is the short-run equilibrium unemployment rate in the presence of errors in price expectations. From the Friedman and Lucas analyses, we have,

$$(u^* - u^n) = g(\pi - \pi^e) \tag{91}$$

Therefore,

$$u = u^n + g(\pi - \pi^e) + [(u - u^*)|\Delta Y/\Delta P] \tag{92}$$

The term $[(u - u^*)|\Delta Y/\Delta P)]$ means that $(u - u^*)$ is conditional on changes in real aggregate demand. If the price level changes do fully reflect the change in nominal aggregate demand, real aggregate demand will not change, so that $[(u - u^*)|\Delta Y/\Delta P]$ will be zero. But if real aggregate demand does change, the unemployment rate will change. To illustrate, the impact of an expansionary monetary policy on aggregate demand, which does not proceed through a change in the price level, will change the actual unemployment rate. The new Keynesian models provide reasons why the third term on the right-hand side of (90) need not be zero.

The implication of the second term on the right-hand side of (90) for the simple Phillips curve (stated as $u = f(\pi)$) is that the Phillips curve for a period t will shift if the inflation rate expected for period t changes. While the new Keynesian models do incorporate the effect of future inflation on the relationship between current inflation and current output, they do not incorporate the effect of errors between the inflation rate in period t and the inflation rate

expected for period *t* itself. While the classical economics models of Lucas and Sargent and Wallace incorporate this element, they do not incorporate the third term on the right-hand side of (90). We conclude that the appropriate Phillips curve needs to incorporate elements of both the modern classical analysis and the Keynesian analysis, with plenty of possibility of revisions, corrections and new theories to fill in the gaps.

17.12 Hysteresis in long-run output and employment functions

The new Keynesian and the modern classical approaches agree that long-run employment and output are independent of inflation and aggregate demand, which also makes them independent of the short-run performance of the economy. This implies the absence of hysteresis, which is broadly defined as the impact of the short run on the long-run performance of the economy. While this absence is analytically taken for granted and is hard to establish empirically, there is some evidence that long-run, or at least long-term, employment depends on the duration and extent of booms and recessions in the economy, and hence on short-run aggregate demand factors.[20]

Conclusions

There are considerable disputes on the underlying macroeconomic model of the economy. The classical propositions of the Friedman–Lucas supply rule propose a relationship in which output is invariant to anticipated inflation and anticipated money supply, while it may not be invariant to the unanticipated values of these variables. There is considerable evidence by now that anticipated monetary policy is not neutral, at least in the short run. This rejection of the central doctrines of the modern classical school opened the way in the 1990s for the resurgence of Keynesian approaches, culminating in the new Keynesian model. This model follows the agenda of the modern classical school that macroeconomic theory has to be based on microeconomic foundations with rational expectations. The core parts of this model are its forward-looking price adjustment equation and the Taylor rule for interest rate setting. Unfortunately, both of these have been strongly rejected in favor of their backward-looking versions.

This chapter has also suggested that the equations for output and unemployment need to encompass several sources of deviations from full employment. One of these sources is errors in expectations, incorporated in the modern classical models, but there are also many other such sources, some of which are incorporated in the Keynesian and new Keynesian models. Further, in addition to the fact that the impact of monetary policy is not neutral, a considerable part of its impact on output and unemployment does not go through price level changes.

Assessment on the choice of the variable (money supply or interest rate) that should be exogenously set by the central bank:

> We didn't abandon the monetary aggregates, they abandoned us.
>
> (Gerald Bouey, Governor of the Bank of Canada in the late 1980s).

20 For instance, Mankiw (2001) suggests a relationship in which the long-run unemployment is a function of the actual short-term unemployment rate.

The profession's assessment on the neutrality of money and discretionary monetary policy:

In their summing up of the effect of monetary changes on output, Milton Friedman and Anna Schwartz (1963) wrote the following:

On the non-neutrality of money:

> Three counterparts of such crucial experiments [of physical science] stand out in the monetary record since the establishment of the Federal Reserve System. On three occasions, the System deliberately took policy steps of major magnitude which cannot be regarded as necessary or inevitable economic consequences of contemporary changes in money income and prices. Like the crucial experiments of the physical scientist, the results are so consistent and sharp as to leave little doubt about their interpretation. The dates are January–June 1920, October 1931, and July 1936–January 1937. These were three occasions … when the Reserve System engaged in acts of commission that were sharply restrictive … each was followed by sharp contractions in industrial production … declines within a twelve-month period of 30 per cent (1920), 24 per cent (1931), and 34 percent per cent (1937), respectively.
>
> (Friedman and Schwartz, 1963, pp. 688–9).

The magnitude of the real effects in these examples is remarkable. Lest it be thought that only monetary contractions have real effects, Friedman and Schwartz (1963, p. 690) also cited three very significant episodes where monetary expansions by the Federal Reserve System caused large increases in industrial production.

Laurence Ball and N. Gregory Mankiw (1994), two of the foremost new Keynesians, write:

On the non-neutrality of money:

> We believe that monetary policy affects real activity. The main reason for our belief is the evidence of history, especially the numerous episodes in which monetary contractions appear to cause recessions. … monetary contractions are a major source of U.S. business cycle.
>
> (Ball and Mankiw, 1994, pp. 128–9).

On the nineteenth-century classical and Friedman's views on the neutrality of money:

> The [pre-Keynes] classical economists never suggested that money was neutral in the short run [but did offer] the key insight … that money is neutral in the long run.
>
> (Ball and Mankiw, 1994, p. 132).

On the sources of the non-neutrality of money:

> We believe that price stickiness is the best explanation for monetary non-neutrality… many prices change infrequently… (though) many (other) prices in the economy are quite flexible. … Other economists, however, accept monetary non-neutrality but resist the assumption of sticky prices. They have been led to develop models of non-neutrality with flexible prices.
>
> (Ball and Mankiw, 1994, pp. 131, 134–5).

On the influence of monetary policy and the role of the central banks, Ball and Mankiw assert that:

> The Fed is a powerful force for controlling the economy…. Policymakers and the press believe that monetary policy can speed up or slow down real economic activity.
>
> (Ball and Mankiw, 1994, 132–3).

On the new Keynesian model

The new Keynesian model is the latest of the macroeconomic models to appear on the scene. It differs from the earlier Keynesian approaches by explicitly deriving its components from microeconomic, intertemporal foundations and rational expectations. In this, it follows the pattern of the modern classical models but differs from the latter by its addition of market imperfections and price stickiness. As Mankiw (2001) points out:

> [Since the time of David Hume, it has been well known that] a monetary injection first increases output and inflation, and later increases the price level (p. C46).
>
> The so-called "new Keynesian Phillips curve" is appealing from a theoretical standpoint, but it is ultimately a failure. It is not at all consistent with the standard dynamic effects of monetary policy, according to which monetary shocks have a delayed and gradual effect on inflation. We can explain these facts with traditional backward-looking (original Phillips curve) models of inflation–unemployment dynamics, but these models lack any foundation in the microeconomic theories of price adjustment.
>
> (Mankiw, 2001, p. C52).

Further, the new Keynesian derivation of the monetary policy rule in the form of a forward-looking Taylor rule does not fare any better when compared with a backward-looking Taylor rule.

How do the above views of Keynesians compare with those of Lucas (1994), who is associated with the modern classical school and has been a major contributor to it? Some of his views are:

On the variant neutrality and non-neutrality of money:

> Sometimes, as in the U.S. Great Depression, reductions in money growth seem to have large effects on production and employment. Other times, as in the ends of the post-World War I European hyperinflations, large reductions in money growth seem to have been neutral, or nearly so. Observations like these seem to imply that a theoretical framework such as the Keynes–Hicks–Modigliani IS/LM model, in which a single multiplier is applied to all money movements regardless of their source or predictability, is inadequate for practical purposes.
>
> (Lucas, 1994, p. 153).

On the lack of adequate knowledge and absence of theories of the variant non-neutrality of money:

> Little can be said to be firmly established about the importance and nature of the real effects of monetary instability, at least for the U.S. in the postwar period. Though it is widely agreed that we need economic theories that capture the non-neutral effects of

> money in an accurate and operational way, none of the many available candidates is without serious difficulties. (p. 153).
>
> Macroeconomic models with realistic kinds of monetary non-neutralities do not yet exist (1994, pp. 153–54). … anticipated and unanticipated changes in money growth have very different effects (1996, p. 679).
>
> [However, on the models that attribute this non-neutrality to unanticipated or random changes in the price level, the evidence shows that] only small fractions of output variability can be accounted for by unexpected price movements. Though the evidence seems to show that monetary surprises have real effects, they do not seem to be transmitted through price increases, as in Lucas (1972).
>
> (Lucas, 1996, p. 679).

In the "Nobel Lecture" (1996), given on his receipt of the Nobel Prize in economics, Robert Lucas added that:

> In summary, the prediction that prices respond proportionately to changes in the long run, deduced by Hume in 1752 (and by many other theorists, by many different routes, since), has received ample – I would say decisive – confirmation, in data from many times and places. The observation that money changes induce output changes in the same direction receives confirmation in some data sets but is hard to see in others. Large-scale reductions in money growth can be associated with large-scale depressions or, if carried out in the form of a credible reform, with no depression at all.
>
> (Lucas, 1996, p. 668).

A central banker's opinions on the neutrality of money, lags and the art of monetary policy

What do central bankers, who are the real-world practitioners of monetary policy, in fact believe and do? In 1997, the central banks of both Canada and the USA had their declared objective as that of aggressively promoting price stability. Both used interest rates, rather than monetary aggregates, as their target and instrument variables. On the dynamics of their policies, a speech by the Governor of the Bank of Canada on October 7, 1997, expressed the stance of the Bank's policy as:

> Too much monetary stimulus can lead to an exhilarating temporary burst of economic activity. But it will almost certainly also lead to inflation-related distortions that undermine both the expansion and the economy's efficiency over the longer term. The end-result, as we know only too well from past experience, is high interest rates, punishing debt loads, recession, and higher unemployment.
>
> A further complication is that it takes between a year to a year and a half for the economy to fully respond to changes in the degree of monetary stimulus … you want to look way ahead to see what is coming, and you want to take action early … That is why monetary policy must focus on future, rather than the present, and why the Bank must act in a forward-looking pre-emptive manner.
>
> (Gordon Thiessen, Governor of the Bank of Canada, 1997)[21].

21 "Challenges ahead for monetary policy." *Remarks to the Vancouver Board of Trade*, 7 October 1997.

Collectively, these assessments of economists from the Keynesian and classical traditions and of a central banker show a high degree of agreement that, in the short run, money can be non-neutral, and more likely to be non-neutral than neutral. However, ignoring the possibility of hysteresis, there is also broad agreement and substantial evidence that long-run output growth is independent of money supply growth.[22] While these conclusions indicate much less divergence on the nature of the economy than the formal models of the different schools convey, there is little agreement on the sources and extent of the potential short-run non-neutrality. There can be several possible sources of non-neutrality, as discussed in the first part of these conclusions, and the reasons for and the extent of non-neutrality can differ at different times and in different countries.

Summary of critical conclusions

- For many economies, the central bank's control of the interest rate as the variable of monetary policy provides a better tool for management of aggregate demand in the economy.
- The LSW model, incorporating the Friedman–Lucas supply function and rational expectations, is often used as the compact form of the modern classical model for short-run macroeconomics. Its policy recommendation is that the central bank should not use changes in the money supply to attempt changes in output and unemployment in the economy.
- While a negative relationship between the rate of unemployment and the rate of inflation seemingly occurs in the LSW model, the coefficients of such a relationship are not invariant to shifts in monetary policy. The Lucas critique would apply to such a relationship.
- The LSW model modified by the Keynesian supply function, as well as the new Keynesian models, allows systematic monetary policy to change output and unemployment in the economy.
- It is by now well established that monetary policy is not neutral in the short run.
- However, a great deal of controversy remains on the reasons for the short-run non-neutrality of monetary policy.

Review and discussion questions

1. What evidence would you need to establish whether or not money supply changes have been the main cause of changes in nominal income? What procedure can you use to determine the direction of causality between the changes in money and in income?
2. Specify the hypotheses on the natural rate of unemployment and the rational expectations. Discuss the logical and historical relationship between them.

   Discuss, for each concept, whether disequilibrium in the economy is consistent with it or not. If it is, discuss the role and usefulness of monetary policy in this state.
3. Discuss the evolution of the 1970s Monetarists' claim that "only monetary policy is effective" to the doctrines of the modern classical school that "no foreseen monetary policy is effective" and "no systematic monetary policy is effective." Would Friedman have subscribed to any of these propositions?

22 This does not imply that output growth is independent of innovations in the financial sector. For this analysis, see Chapter 24.

4. Inflation and unemployment are two crucial economic items of interest to the public. Is it possible to explain one without the other? Present at least one theory that explains each independently of the other and one theory that establishes their interdependence. Is there a genuine difference between these theories or do they merely represent a distinction between the impact and long-term effects of an exogenous change to their determinants?
5. Present a model with rational expectations and the Friedman–Lucas supply function. If policy makers and the public have the same information, can stabilization policies in a stochastic context change aggregate demand and output (i) in the short run, (ii) in the long run?
6. Present a model with rational expectations and the new Keynesian supply function. If policy makers and the public have the same information, can stabilization policies in a stochastic context affect aggregate demand and output (i) in the short run, (ii) in the long run?
7. Why do models with rational expectations have difficulty in explaining the persistence of output from its trend and unemployment from the natural rate? What are some of the reasons given for this persistence? If this persistence were incorporated in them, what would be their implications for the effectiveness of monetary policy: could activist monetary policy stabilize output and the unemployment rate? Discuss in the context of a specific model embodying such persistence.
8. Outline the development and current theoretical status of the tradeoff between the unemployment rate and the rate of inflation.
9. "Between 1930 and 1990, macroeconomic theory went through a full circle. The classical view of the 1920s was discarded in the 1930s and 1940s by Keynesian theory, but the latter, in turn, was gradually eroded to the point that the dominant theory by the 1980s had again become a form of the classical one." Discuss.
10. "In the past forty years, macroeconomic theory has come full circle. The Keynesian views of the economy were discarded in the mid-1970s by the resurgent classical theory, but the latter, in turn, has been eroded to the point that the dominant theory is again a form of the Keynesian one." Discuss.
11. Modern classical macroeconomics argues that anticipated monetary policy does not have real effects. The new classical macroeconomics argues that anticipated fiscal policy also does not have real effects. Adapt the Lucas–Sargent–Wallace model to explicitly incorporate both of these propositions. From this model, what would you conclude for the effects of (i) an anticipated bond-financed deficit, (ii) an anticipated money-financed deficit? Specify your procedure and estimating equations for testing the validity of your conclusions.
12. Consider an economy with the following structure:
    Aggregate demand:

$$y_t = M_t - P_t + \mu_t \qquad \text{(A quantity theory type equation)}$$

    Aggregate supply:

$$y_t = y^{\mathrm{f}}{}_t + \gamma(P_t - P^{\mathrm{e}}{}_t) + \eta_t$$

    where the symbols have the usual meanings and are in logs. μ and η are random errors.

Expectations are formed in one of the following alternative ways:

(a) Rational expectations:

$$P^{e}_{t} = E_{t-1}P_t$$

(b) Adaptive expectations:

$$P^{e}_{t} - P^{e}_{t-1} = (1-\lambda)(P_{t-1} - P^{e}_{t-1})$$

(i) Given rational expectations, solve for P_t and y_t (in terms of the money supply and random errors, etc.).
(ii) Given adaptive expectations, solve for P_t and y_t (in terms of the money supply, etc.).
(iii) For the two expectations hypotheses, derive the time patterns of the response of the price level to a unit change in the money supply.
(iv) For the two expectations hypotheses, derive the time pattern of the response of real output to a unit change in the money supply.
(v) Discuss the differences implied by the two hypotheses for the impact of monetary policy on real output and prices.

13. Consider the following model:
Aggregate supply:

$$y_t = \gamma(P_t - E_{t-1}P_t) + \gamma(P_t - E_{t-2}P_t)$$

Aggregate demand:

$$y_t = M_t - P_t + \mu_t$$
$$\mu_t = \mu_{t-1} + \eta_t$$

where y, P and M have the usual meanings and are in logs. η is a serially uncorrelated error with mean zero and variance σ^2.

(i) How would you justify the above aggregate supply function and how does it differ from the Lucas one?
(ii) Suppose that the central bank can observe μ_{t-1} but not μ_t when it sets the money supply. Is there then a role for systematic monetary policy?
(iii) Given (ii), suppose that the central bank wants to set the money supply to minimize the variance $E_{t-1}(y_t - y^{f})^2$ of output about its full-employment level? What monetary policy would it follow?
(iv) If the policy in (iii) has always been followed, is there any way of using econometric evidence to differentiate between the pattern shown by this economy and one in which the economy had a Lucas supply function?

14. Suppose the economy is described by:
Aggregate supply:

$$y_t = y^{f}$$

Aggregate demand and fiscal policy:

$$y_t = \alpha_0 + \alpha_1(M_t - P_t) + \alpha_2 E_{t-1}(P_{t+1} - P_t) + \alpha_3 z_t + \mu_t \qquad \alpha_1, \alpha_2, \alpha_3 > 0$$

$$z_t = \gamma_0 + \gamma_1 z_{t-1} + \eta_t$$

Money supply:

$$M_t = M_0 + \nu_t$$

where all variables are in logs. μ, η and ν are random disturbance terms. y, M and P have the usual meanings, and z is a real fiscal variable.

Find the equilibrium solutions for y_t and P_t. How do systematic or anticipated changes in the nominal money supply M_0 affect y_t and P_t? How would unanticipated changes in the nominal money supply M_t affect y_t and P_t? How would anticipated and unanticipated changes in the fiscal variable affect y_t and P_t?

15. Suppose the output supply in the model of question 14 is not the full-employment level but is determined by demand. Re-do the answers to question 14.
16. Suppose the supply function in the model of this question 14 is changed to:

$$y_t = y^{\mathrm{f}} + \gamma(P_t - E_{t-1}P_t) \quad \gamma > 0$$

Re-do the answers to question 14.

17. Suppose the price level in question 14 is fixed at $\underline{P}$ (making the model a fixed price one, compared with the preceding flexible price model). Re-do the answers to this question.
18. Specify the three types (backward-looking, contemporaneous and forward-looking) of the Taylor rule on interest rates, and discuss their validity. Why does the forward-looking form implied by the new Keynesian model seem to do badly relative to the others, and especially relative to the backward-looking one?
19. Write the general dynamic form of the sticky-price new Keynesian Phillips curve as:

$$\pi_t = \beta E_t \pi_{t+1} + \alpha(u_t - u^{\mathrm{n}}) + v_t$$

where π is the inflation rate, u is the unemployment rate, u^{n} is the natural rate, $v_t = \rho v_{t-1} + \eta_t$ and η is a random variable with a zero mean and constant variance. Mankiw (2001) argues that this relationship is not valid and that the general form of the valid relationship is backward-looking in inflation. Its corresponding form would then be:

$$\pi_t = \beta \pi_{t-1} + \alpha(u_t - u^{\mathrm{n}}) + v_t$$

Is this form consistent with the original Phillips curve? Provide the theoretical justification for it offered by the sticky information model, and critically evaluate this contribution.

20. Given the generally poor performance of the forward-looking forms of the Taylor rule and the Phillips curve implied by the NK model, especially relative to their backward-looking simpler versions, discuss whether the new Keynesian reformulation of Keynesianism has proved to be a failure. If this is so, are there any elements of the NK ideas that should be preserved in further research?

References

Ball, L. "Policy rules for open economies." In J.B. Taylor, ed., *Monetary Policy Rules*. Chicago: Chicago University Press, 1999.

Ball, L. "Policy rules and external shocks." *NBER Working Paper* no. 7910, 2000.

Ball, L., and Mankiw, N.G. "A sticky-price manifesto." *Carnegie-Rochester Series on Public Policy*, 41, 1994, pp. 127–51.

Barro, R.J. "Unanticipated money growth and unemployment in the United States." *American Economic Review*, 67, 1977, pp. 101–16.

Bernanke, B.S., and Woodford, M. "Inflation forecasts and monetary policy." *Journal of Money, Credit and Banking*, 29, 1997, pp. 653–84.

Bullard, J., and Keating, J.W. "The long-run relationship between inflation and output in postwar economies." *Journal of Monetary Economics*, 36, 1995, pp. 477–96.

Chu, J., and Ratti, R.A. "Effects of unanticipated monetary policy on aggregate Japanese output: the role of positive and negative shocks." *Canadian Journal of Economics*, 30, 1997, pp. 722–41.

Clarida, R., Gali, J. and Gertler, M. "The science of monetary policy: a new Keynesian perspective." *Journal of Economic Literature*, 37, 1999, pp. 1661–707.

Cover, J.P. "Asymmetric effects of positive and negative money shocks." *Quarterly Journal of Economics*, 107, 1992, pp. 1261–82.

Depalo, D. "Japan: the case for a Taylor rule? A simple approach." *Pacific Economic Review*, 11, 2006, pp. 527–46.

Friedman, M., and Schwartz, A.J. *A Monetary History of the United States, 1867–1960*. Princeton, NJ: Princeton University Press, 1963.

Frydman, R., and Rappoport, P. "Is the distinction between anticipated and unanticipated growth relevant in explaining aggregate output?" *American Economic Review*, 77, 1987, pp. 693–703.

Gali, J. "How well does the IS–LM model fit post-war U.S. data?" *Quarterly Journal of Economics*, 107, 1992, pp. 709–38.

Gali, J. and Gertler, M. "Inflation dynamics: a structural econometric analysis." *Journal of Monetary Economics*, 44, 1999, pp. 195–222.

Levin, A.T., Wieland, T.V. and Williams, J.C. "Robustness of simple monetary policy rules under model uncertainty." In J.B. Taylor, ed., *Monetary Policy Rules*. Chicago: University of Chicago, 1999, pp. 263–99.

Levin, A.T., Wieland, T.V. and Williams, J.C. "The performance of forecast-based monetary policy rules under model uncertainty." *Federal Reserve System Working Paper* no. 2001-39, 2001.

Lucas, R.E., Jr. "Expectations and the neutrality of money." *Journal of Economic Theory*, 4, 1972, pp. 103–24.

Lucas, R.E., Jr. "Some international evidence on output–inflation tradeoffs." *American Economic Review*, 63, 1973, pp. 326–34.

Lucas, R.E., Jr. "Comments on Ball and Mankiw." *Carnegie-Rochester Series on Public Policy*, 41, 1994, pp. 153–5.

Lucas, R.E., Jr. "Nobel lecture: monetary neutrality." *Journal of Political Economy*, 104, 1996, pp. 661–82.

Mankiw, N.G. "The inexorable and mysterious tradeoff between inflation and unemployment." *Economic Journal*, 111, 2001, pp. C45–C61.

Mankiw, N.G., and Reis, R. "Sticky information versus sticky prices: a proposal to replace the new Keynesian Phillips curve." *Quarterly Journal of Economics*, 117, 2002, pp. 1295–328.

Mankiw, N.G., and Reis, R. " Pervasive stickiness." *American Economic Review*, 96, 2006a, pp. 164–9.

Mankiw, N.G., and Reis, R. " Sticky information in general equilibrium." *NBER Working Paper* no. 12605, 2006b.

Maria-Dolores, R., and Vazquez, J. "How does the new Keynesian monetary model fit in the U.S. and the Eurozene? An indirect inference approach." *Topics in Macroeconomics*, 6, 2006, article 9, pp. 1–49.

Mishkin, F.S. "Does anticipated aggregate demand policy matter? Some further econometric results." *American Economic Review*, 72, 1982, pp. 788–802.

Mosser, P. "Changes in monetary policy effectiveness: evidence from large macroeconomic models." *Federal Reserve Board of New York Quarterly Review*, 17, 1992, pp. 36–51.

Rudd, J., and Whelan, K. "Can rational expectations sticky price models explain inflation dynamics?" At www.federalreserve.gov/pubs/feds/2003/200346, 2003.

Sargent, T.J., and Wallace, N. "Rational expectations and the theory of economic policy." *Journal of Monetary Economics*, 2, 1976, pp. 169–83.

Wang, S., and Handa, J. "Monetary policy rules under a fixed exchange rate regime: empirical evidence from China." *Applied Financial Economics*, 17, 2007, pp. 941–50.

18 Walras's law and the interaction among markets

This chapter focuses on Walras's law, which is fundamental to macroeconomic analysis. It underlies the IS–LM model, which has four goods – commodities, money, bonds and labor – but has explicit analysis of only three of them. The foundations of Walras's law and its validity in disequilibrium are rigorously examined in this chapter. This is also done for Say's law. The analyses of Walras's and Say's laws are followed by the derivation of their joint implications for the dichotomy between the real and the monetary sectors, and the neutrality of money.

The Pigou and real balance effects concern the impact of changes, brought about by changes in the price level, in the value of financial assets and of real balances, respectively, on the demand for commodities. They played a key role in doctrinal disputes on whether an economy functioning below full employment would return to full employment. This chapter examines their nature and empirical relevance.

These discussions are followed by analyses of effective Clower and Drèze demand and supply functions versus notional ones.

Key concepts introduced in this chapter

- A law versus an equilibrium condition
- Walras's law
- Say's law
- Dichotomy between real and monetary sectors
- Pigou effect
- Real balance effect
- Notional demand and supply functions
- Clower effective demand and supply functions
- Drèze functions

There are very few propositions in economics that have been considered to be so far beyond dispute as to be labeled "laws." Among these are Walras's law and Say's law. The former represents a constraint on all goods in the economy while the latter represents a constraint on the commodity market alone. This chapter will consider these laws for the closed economy, though the arguments can be easily extended to the open economy.

A statement in economics worthy of being called a *law* must be more than a behavioral relationship or an equilibrium condition – both of which are not necessarily valid at a given time for a given economy – since, otherwise, there would not be a special reason

for assigning it a distinctive term with the compelling connotation that the word "law" possesses. Hence, while there can be several ways of defining what can be designated as a law in economics, we will define it as a statement that holds without exception under any and all conditions, so that it is an identity – or, if one is willing to be more tolerant, it must be a statement that approximates this degree of applicability. There are very few statements of this nature in microeconomics, let alone in macro or monetary economics.[1] The classical paradigm asserts that Walras's law is one of this extremely select group. However, as we discuss in Section 18.8 below, there are serious and well-founded arguments against its being a law or an identity. In particular, while it holds in the general equilibrium states of the economy, its implications for the dynamic analysis of prices, wages and interest rates need not be always valid. Hence, we conclude that it is not an identity and, therefore, not a law.

Say's law asserts that the supply of commodities creates an equal demand for them. This chapter shows that it does not hold as an identity in the modern economy with financial assets, so that it is clearly not a law for the modern economy. It really should not be a part of any modern macroeconomic analysis.

Sections 18.1 to 18.3 present the derivation of Walras's law and its implications. Section 18.4 deals with Say's law and finds it inappropriate for monetary economies. Section 18.5 discusses the implications of the joint assumption of Walras's law and Say's law for the neutrality of money and the dichotomy between the real and monetary sectors. Sections 18.6 and 18.7 present the wealth and real balance effects. Both these concepts have already been presented in Chapter 3 in the context of the Walrasian general equilibrium model of the economy. Sections 18.8 to 18.12 examine the conditions for the breakdown of Walras's law and the implications of this breakdown.

18.1 Walras's law

Walras's law is the statement that for any economy, over any given period of time, the sum of the market *values* of all the goods demanded must equal the sum of the market values of all the goods supplied. For the closed economy,[2] we define "goods" in this chapter, as in earlier chapters, to refer to commodities, money, labor (or leisure) and non-monetary financial assets ("bonds").

To explain Walras's law intuitively for a pure exchange economy, first start with the constraint on the individual's demands for goods. Assume that the individual initially possesses some commodities, some money and some bonds,[3] and that their total nominal value at current market prices is his nominal wealth ψ. He will also want to supply some labor, with a nominal labor income at current wage rates being equal to Y. Assuming that he spends this amount to acquire the commodities, money and bonds that he wants to hold

1 Another proposition that is sometimes referred to as a law is the "law of one price," which in international trade translates to absolute purchasing power parity between countries. However, it is often rejected by empirical evidence, so that it cannot be properly regarded as a law.

Another proposition often labelled as a law is that "the demand curves for commodities slope downwards." However, this is violated for many goods, such as those with snob appeal, which have upward-sloping demand curves. Therefore, even this proposition is not an identity and not really a law.

2 The open economy has an additional good in the form of foreign exchange held by the private and public sectors.

3 Bond holdings can be positive (making the individual a net lender) or negative (making the individual a net borrower).

or use in the current period,[4] the total value ψ of his initial holdings of goods plus his labor income Y must equal his total expenditures on commodities, money and bonds. Since the individual's total expenditures on his purchases of commodities, money and bonds must sum to $[\psi + Y]$, the demand for any one of these three goods can be derived by subtracting his expenditures for the other two goods from $[\psi + Y]$.

If we now sum over all the individuals in the economy, the aggregate initial holdings of goods (including labor) plus the current national output becomes their supply and the aggregate expenditures on them become the value of the quantities demanded. Further, the total value of all the goods demanded must equal the total value of all the goods supplied. This is Walras's law, and it is the aggregate counterpart of the individual's budget constraint. It is often simplified to the statement that *the supply of all goods in the economy must equal the demand for all goods in the economy*.

Deriving Walras's law for a five-good closed economy

We prove Walras's law for a closed economy with a government sector and for its *five* goods – commodities, money, bonds, equities, and labor.[5] The assumption basic to Walras's law is that there is no "free" disposal of goods (i.e. economic agents do not just throw them away) so that the goods that are produced or inherited from the past are either consumed, demanded for some other reason (such as for carrying to the future as saving in the form of commodities) or exchanged against money or bonds. Assume that the closed economy has three economic agents – households, firms and the government (including the central bank) – so that there are several budget constraints to consider in the analysis. The definitions of the symbols in the constraints are given after the specification of the constraints, which are as follows.

HOUSEHOLDS' BUDGET CONSTRAINT

This constraint specifies that the payments by households for all goods (commodities, money, bonds and equities) bought must equal the funds available from the initial endowments (i.e. inherited stocks) of goods, labor income and the profits received from firms as distributed profits. That is:

$$p_c c^{\mathrm{dh}} + p_c m^{\mathrm{dh}} + p_b b^{\mathrm{d}} + p_e e^{\mathrm{d}} \equiv p_c \underline{c}^{\mathrm{s}} + \underline{M}^{\mathrm{h}} + p_b \underline{b}^{\mathrm{s}} + p_e \underline{e}^{\mathrm{s}} + W n^{\mathrm{s}} + \pi^{\mathrm{dis}} \tag{1}$$

FIRMS' BUDGET CONSTRAINTS

Firms face three constraints. The first one specifies that the funds available to the firms from the sales of their products (to households for consumption, to other firms for investment and to government) and of new equities must equal the sum of the payments to the labor employed and funds used for the firm's investments, money holdings or distributed profits. The firms' second constraint specifies that total profits can be either distributed or retained/undistributed

4 We have assumed that the individual is a supplier of labor services and firms are the buyers of such services.

5 We have included three financial goods, money, bonds and equities, in the following analysis in order to show that the separate treatment of equities does not destroy Walras's law. Note that in the rest of this book and in monetary economics generally, "bonds" are defined to include all non-monetary assets and therefore would also include equities.

by the firm. The third constraint specifies that the firm's investment must be financed from its retained earnings and the issue of new equities.

$$p_c(c^{\mathrm{sh}} + i + g) + p_e(e^{\mathrm{s}} - \underline{e}^{\mathrm{s}}) + \underline{M}^{\mathrm{f}} \equiv Wn^{\mathrm{d}} + p_c i + p_c m^{\mathrm{df}} + \pi^{\mathrm{dis}} \tag{2}$$

$$\pi \equiv \pi^{\mathrm{dis}} + \pi^{\mathrm{undis}} \tag{3}$$

$$p_c i \equiv \pi^{\mathrm{undis}} + p_e(e^{\mathrm{s}} - \underline{e}^{\mathrm{s}}) \tag{4}$$

GOVERNMENT'S BUDGET CONSTRAINT

This constraint specifies that the government must finance its deficits or surpluses by the issue of new money and bonds.

$$p_c(g - t) \equiv (M^{\mathrm{s}} - \underline{M}^{\mathrm{s}}) + p_b(b^{\mathrm{s}} - \underline{b}^{\mathrm{s}}) \tag{5}$$

The definitions of the symbols in the preceding five identities are:

p_c	price of commodities	$\underline{M}$	existing nominal money stock (held by households and firms)
p_b	price of government bonds	$\underline{b}$	existing stock of bonds (issued by the government, held by households)
p_e	price of equities	$\underline{e}$	existing stock of equities (issued by firms, held by households)
W	nominal wage rate = rental price of labor	g	real government expenditures on commodities
c^{sh}	supply of commodities to households	t	government's tax revenues
c^{sf}	total supply of commodities by firms ($\equiv c^{\mathrm{sh}} + i + g$)	i	real investment by firms
m	real money balances	π	total nominal profits of firms
M	nominal money stock	π^{dis}	nominal distributed profits
b	quantity of bonds	π^{undis}	nominal retained (undistributed) profits of firms
e	quantity of equities		
n	number of workers (labor)		
$\underline{c}$	existing stocks of commodities (held by households)		

The superscripts d and s stand for demand and supply respectively. The superscripts h and f stand for households and firms respectively. Underlining indicates that the value of the variable is given exogenously or was inherited from the past. Note that $c^{\mathrm{d}} \equiv c^{\mathrm{dh}} + i + g$, $c^{\mathrm{s}} \equiv c^{\mathrm{sf}} + \underline{c}^{\mathrm{s}}$ and $\underline{M} \equiv \underline{M}^{\mathrm{h}} + \underline{M}^{\mathrm{f}}$. For simplification, (1) to (5) make the usual assumption that only firms issue equities and that only the government issues bonds.

Note that (1) to (5) are all *identities*, designated by the use of the symbol $\equiv$, and result from the assumption that nothing is just thrown away by economic agents since doing so would be irrational. They imply that:

$$p_c(c^{\mathrm{d}} - c^{\mathrm{s}}) + (M^{\mathrm{d}} - M^{\mathrm{s}}) + p_b(b^{\mathrm{d}} - b^{\mathrm{s}}) + p_{\mathrm{e}}(e^{\mathrm{d}} - e^{\mathrm{s}}) + W(n^{\mathrm{d}} - n^{\mathrm{s}}) \equiv 0 \tag{6}$$

where the left-hand side is the sum of the nominal excess demands for commodities, money, bonds, equities and labor. (6) is one of the ways of stating Walras's law: *the sum of the*

nominal excess demands for all goods in the economy must be zero. (6) restated in terms of excess demands is:

$$E_c{}^{\mathrm{d}} + E_m{}^{\mathrm{d}} + E_b{}^{\mathrm{d}} + E_e{}^{\mathrm{d}} + E_n{}^{\mathrm{d}} \equiv 0 \tag{7}$$

where $E_k{}^{\mathrm{d}}$ is the excess *nominal* demand for the kth good, $k = c, m, b, e, n$.

The distinction between Walras's law and Walrasian general equilibrium models

Walrasian general equilibrium models assume that equilibrium exists in all markets and that markets are perfect, i.e. perfectly competitive and efficient, so that they always clear. Walras's law does not embody this assumption of perfect markets, or that that any or all markets are in equilibrium. Further, Walrasian general equilibrium models make statements about the equilibrium state, which is not an identity, while Walras's law *is* an identity. Therefore, the two concepts of Walras's law and Walrasian general equilibrium models are quite different. However, Walras's law is a requirement of all Walrasian general equilibrium models, as of other models of the economy, whereas Walrasian general equilibrium is not a requirement of Walras's law.

18.1.1 Walras's law in a macroeconomic model with four goods

We have differentiated between bonds and equities to capture the financial structure of the economy in a more realistic manner and to show that Walras's law will hold for this structure. In the general case, it will also hold for any economy no matter how its goods are categorized. Since macroeconomic theory usually treats firms' equities and government bonds as the composite good "bonds" (see Chapter 13), we will at this stage shorten (6) in line with this usage. Walras's law for this more compact *four-good* closed economy can be stated in the following two alternate ways.

(i) $$E_c{}^{\mathrm{d}} + E_m{}^{\mathrm{d}} + E_b{}^{\mathrm{d}} + E_n{}^{\mathrm{d}} \equiv 0 \tag{8}$$

where b (bonds) now represents all non-monetary financial assets in the economy. (8) is the statement that *the sum of the excess demands for the four goods is identically zero*.

(ii) Equation (8) can be rewritten as:

$$p_{\mathrm{c}}c^{\mathrm{d}} + M^{\mathrm{d}} + p_b b^{\mathrm{d}} + Wn^{\mathrm{d}} \equiv p_c c^{\mathrm{s}} + M^{\mathrm{s}} + p_b b^{\mathrm{s}} + Wn^{\mathrm{s}} \tag{9}$$

which is the statement that *the sum of the nominal demands for all goods identically equals the sum of the nominal supplies of all goods in the economy*.

For the general case of K markets, (8) and (9) generalize to:

$$\sum_{k=1}^{K} E_k^{\mathrm{d}} \equiv 0 \tag{10}$$

$$\sum_{k=1}^{K} x_k^{\mathrm{d}} = \sum_{k=1}^{K} x_k^{\mathrm{s}} \tag{10′}$$

where x_k is the quantity of the kth good and there are K goods in the economy. (10) or (10′) is the general statement of Walras's law.

Adjustment costs, which slow the adjustment in the demands and/or supplies of goods so that their short-run values differ from their long-run ones, do not change the derivation or applicability of Walras's law, nor does the introduction of uncertainty and rational expectations into the model. However, the validity of Walras's law for dynamic analysis does become questionable if the labor and commodity markets do not clear on a continuous basis. Sections 18.8 to 18.10 will examine this issue.

18.1.2 The implication of Walras's law for a specific market

Walras's law by itself does not assert equilibrium in each market or in any one specific market, but is a statement covering all markets in the economy. (10) implies that:

$$E_K{}^{\mathrm{d}} \equiv -\sum_{k=1}^{K-1} E_k^{\mathrm{d}} \tag{11}$$

where $E_K{}^{\mathrm{d}}$ is excess demand in the Kth market. This condition asserts that the excess nominal demand in the Kth market equals the sum of the excess nominal demands in the other $(K-1)$ markets, where we can arbitrarily designate the market for any specific good as the Kth market. (11) implies that:

$$\text{If } E_k{}^{\mathrm{d}} = 0, \text{ for } k = 1,\ldots,K-1, \text{ then } E_K{}^{\mathrm{d}} = 0 \tag{12}$$

That is, if there exists equilibrium in $K-1$ markets, then there would also be equilibrium in the Kth market. (12) is also sometimes used as a way of stating Walras's law. In general equilibrium analysis, (11) and (12) allow the analysis to dispense with the explicit treatment of one of the markets in the economy. Note, however, that such an omitted market continues to exist and to function but its treatment is pushed into the implicit state. Further, the solution of any of the three markets for the three prices (P, p_b, W) in (9) would be identical, irrespective of which market is omitted from the explicit analysis. Furthermore, the general equilibrium values of the real variables, such as output, employment, consumption, etc., will also be identical.

18.2 Walras's law and selection among the markets for a model

Implications of Walras's law for general equilibrium analysis

As shown above, Walras's law permits the general equilibrium conditions for any three out of the four goods in the closed-economy macroeconomic model to be explicitly specified for the solution of the overall equilibrium of the economy. Therefore, the complete model can explicitly set out the equations for any of the following four groups of markets:

(I) commodities, money and labor markets;
(II) bond, money and labor markets;

(III) commodities, bond and labor markets;
(IV) commodities, money and bond markets;

Grouping I was used by Keynes, and has become the standard set used in modern macroeconomics. The neoclassical and the modern classical models also follow this pattern. As is discussed in greater detail in Chapter 19, the traditional classical economists, prior to Keynes, had generally favored grouping II, with the bond market determining the interest rate in the loanable funds theory, the market for money determining the price level by the quantity theorem, and the labor market determining employment and – through the production function – determining output. Walras's law implies that each of the above sets would provide the same general equilibrium solution of the values of the endogenous variables, even though the explicitly specified three markets differ between the sets.

Implications of Walras's law for the dynamic analysis of markets

While the different approaches may yield the same general equilibrium values of the variables, the general pattern of economic analysis identifies a "price" variable with each market. Thus, the price level is the "price" of commodities, and microeconomic analysis identifies its determination with the demand and supply of commodities. The interest rate is similarly the one-period "price" of loans/bonds, and microeconomic analysis would identify its determination by the market for loans/bonds. The wage rate is the rental price of labor, and microeconomic analysis identifies its determination with the labor market. Hence, there is no "price" left to identify as the price of money, so that there is no price variable that can be uniquely identified with the demand and supply of money. *There can, in fact, be no such unique variable in a monetary economy since money is itself the good in which the prices of other goods (commodities, bonds and labor) are measured.* Thus, $1/P$ is sometimes said to be the value of money(in commodity units) while, at other times, the interest rate is said to be its opportunity cost, and many pre-Keynesian economists designated its value (in labor units) as $1/Pw$.

This reasoning suggests that, if a model is to capture the dynamic movements of prices, wages and the interest rate in real-world markets, it needs to take account of the empirical reality that the price of a particular good responds *in the first instance* to the excess demand only in the market for that good. This price should not respond to the excess demands for other goods, unless these demands spill over into the excess demand for the good in question. Therefore, for dynamic analysis, the appropriate assumptions would be:

$$\partial P_t/\partial t = f(E_{ct}{}^{\mathrm{d}})$$

$$\partial R_t/\partial t = f(E_{bt}{}^{\mathrm{d}})$$

$$\partial W_t/\partial t = f(E_{nt}{}^{\mathrm{d}})$$

where R is the nominal yield on bonds. These dynamic functions seem to be consistent with common intuition and economic folklore on price adjustments in markets. Several studies of the dynamic analysis of the price level and the interest rate have in recent years followed this pattern: that is, making changes in the price level a function of the commodity

market disequilibrium and making interest rate changes a function of the bond market disequilibrium[6]. Hence, for dynamic analysis, the preferred overall macroeconomic model needs to specify the commodity, labor and bond markets, while excluding the market for money balances.[7]

18.3 Walras's law and the assumption of continuous full employment

If it is assumed that the labor market is *continuously* (always) in equilibrium, $n^{\mathrm{d}} = n^{\mathrm{s}} = n^{\mathrm{f}}$ where n^{f} stands for full employment, (8) becomes,

$$E_c{}^{\mathrm{d}} + E_m{}^{\mathrm{d}} + E_b{}^{\mathrm{d}} \equiv 0 \tag{13}$$

The underlying assumption behind (13) is that equilibrium exists on a continuous basis in the labor market – so that $E_n{}^{\mathrm{d}} = 0$, by assumption – but does not do so in the commodities, money and bond markets. In fact, in economies with developed financial markets, the most plausible assumption would be that the money and bonds markets adjust the fastest to clear any disequilibrium. Regarding the commodity and labor markets, the labor markets are the slowest to adjust since they are characterized by long-term explicit or implicit contracts between the firms and their employees. In fact, one of the major disputes dividing economists into the main groupings of Keynesians and modern classical economists is precisely over the issue of whether the labor markets will or will not clear over a reasonably short period, let alone continuously.

Hence, the underlying assumption of (13) that the labor market continuously clears – while the commodity, money and bond markets do not do so – is highly questionable as a basis for macroeconomic analysis. However, while this assumption is of doubtful validity, it is often made, as in the modern classical model specified in Chapter 14.

18.4 Say's law

Say's law is attributed to the writings of Jean-Baptiste Say in the first quarter of the nineteenth century[8] and is considered to be one of the underpinnings of the traditional (pre-Keynesian) classical macroeconomic model. Its usual statement is that "*supply creates its own demand*."[9] Since this statement is meant to refer exclusively to commodities rather than to the other goods in the economy and is meant to apply only in the aggregate over all commodities, it can be more precisely formulated as: *the aggregate supply of commodities creates its own aggregate demand*.

6 See Shaller (1983). Shaller's dynamic analysis of prices and interest rates specifies changes to these in disequilibrium in terms of the commodity and bond market equations, rather than in the commodity and money market equations.

7 This point will be taken up again in Chapter 19 on the determination of the rate of interest.

8 Baumol (1999) claims that the appellation "Say's law" was given in the early twentieth century to ideas of which Say was an enunciator, but that they did not originate with him. Further, Say and other writers espousing these ideas did not make claims to its being a "law."

9 Baumol (1999) attributes this mode of statement to Keynes and states that "Keynes, at best, did not get it quite right." (p. 195). He attributes the interpretation of Say's law as an identity to Oskar Lange (1942).

Say's law was implicit in many of the expositions of the traditional classical model. It can be found in the writings not only of Say but also of Adam Smith[10] in the late eighteenth century, and David Ricardo, John Stuart Mill, Alfred Marshall and others in the nineteenth century. Chapter 1 provided some discussion of Say's law[11] and should be reviewed at this stage. The law's general implication was that there could not exist either an excess demand or an excess supply of commodities in the closed economy. The core reason given for Say's law was that the whole of saving was converted into investment, so that no part of it was converted into money holdings since the former had a positive return while the latter did not.[12]

The statement that the supply of commodities creates its own demand has two components. One is the causality from supply to demand and the other is their identical amount. On the former, the argument runs as follows: the supply of commodities creates income which the recipients must spend on commodities, so that any increase in the aggregate supply of commodities creates a corresponding increase in the aggregate demand for them. This argument is fallacious in the commodities–monetary–bonds economy, since the increase in income could be partly or wholly used to increase money or bond holdings, so that the increase in the aggregate supply of commodities would induce a less than corresponding increase in their aggregate demand. Conversely, if the economic agents choose to increase their commodity demand by running down their money or bond holdings, an increase in the aggregate demand for commodities will come about without a corresponding prior increase in their supply. Hence, the causal argument behind Say's law is not valid in modern economies.

As argued earlier, economics does not use the term "law" to refer to equilibrium conditions, otherwise the equilibrium conditions for the bonds, money and labor markets would also be referred to as laws. These conditions are, in fact, never referred to as laws. Similarly, the interpretation of Say's law as an equilibrium condition does not merit the designation of a "law." Therefore, we will henceforth treat Say's law as an identity. That is, Say's law would be interpreted as: in the aggregate, the demand for commodities *always* equals their supply, with causality taken to run from the latter to the former and not from the former to the latter.

For a rudimentary (barter) economy in which the only traded goods are commodities, Walras's law simplifies to the statement that the aggregate expenditure on commodities must always equal the aggregate income from their sale. That is, for such an economy, Walras's law implies that the demand and supply of commodities are always equal, which is identical with Say's law. Therefore, for such a rudimentary economy, Say's law can be derived as an identity from the aggregate budget constraint for the economy, obtained from summing over the budget constraints of all its economic units. Note that financial assets are excluded by assumption from such an economy, so that substitution between commodities and either of the financial assets (money and bonds) is excluded.

10 For example, Adam Smith wrote that "What is annually saved is as regularly consumed as what is annually spent, and nearly in the same time too … saving … is immediately employed as capital either by himself or some other person. … The consumption is the same but the consumers are different." (Baumol, 1999, p. 200).

11 There are disputes about both the attribution of Say's law and how accurately it reflected the ideas of Say and other writers who expounded it. Our concern in this book is not with the historical attribution of this statement or whether or not it accurately reflects the ideas of Say or his contemporaries, but rather with examining its implications and validity.

12 Chapter 5 on speculative demand and Chapter 6 on the buffer stock of money imply that this reasoning is not valid for monetary economies.

Conversely, Say's law is inapplicable to an economy in which both financial assets and commodities exist and some substitution can occur between them. Hence, Say's law cannot be validly applied to modern economies, all of which have money and bonds among their traded goods.[13]

Some invalid implications of Say's law

Since Say's law is inapplicable to monetary economies, its application to such economies leads to conclusions that are objectionable and inappropriate for monetary economies. The following arguments present some of these.

1 From Say's law, since the commodity sector *always* clears irrespective of the price level of commodities, we have:

$$y^{\mathrm{d}}(P_0) \equiv y^{\mathrm{s}}(P_0)$$

as well as:

$$y^{\mathrm{d}}(\lambda P_0) \equiv y^{\mathrm{s}}(\lambda P_0)$$

for any $\lambda > 0$. Hence, the price level becomes indeterminate in so far as the commodity sector is concerned: at a price P_0, there is equilibrium and zero excess demand in the commodities market, as it is at any price $\lambda \mathrm{P}_0$. Therefore, additional information is necessary to determine the price level.[14]

2 Say's law asserts that the aggregate demand for commodities always equals their aggregate supply, irrespective of the price level. The price level is thus not affected by shifts in the commodity market. For example, an increase in investment, exports, fiscal deficits, etc., causing an increase in aggregate demand, cannot increase the price level, contrary to both intuition and the implications of almost every macroeconomic model.

3 Say's law on its own asserts that the supply and demand for commodities are always identical, irrespective of the interest rate and the level of income in the economy. Hence, in IS–LM models, the IS relationship (curve) would span the whole (r, y) space, rather

13 Some economists define a "weak" form of Say's law as: *in equilibrium, the aggregate demand for commodities equals their aggregate supply*. This confinement of Say's law to an equilibrium condition is hardly very restrictive for the commodity market and merely becomes the specification of the IS relationship in macroeconomic analysis. As only an equilibrium condition, it will allow the possibility of disequilibrium where it will not hold. With disequilibrium in the commodity market, the economy will sometimes, but not at other times, have the aggregate demand for commodities equal to their supply. Interpreted in this way, Say's law would hardly merit the designation of a "law" since we do not give the equilibrium conditions for the money market, the bond market, the labor market or the foreign exchange markets – or for microeconomic markets such as those for apples – the designation of laws.

14 The traditional classical approach followed this procedure and used the quantity theory to specify the price level. To do this, Say's law was supplemented by the equation $M = P \cdot m^{\mathrm{d}} = P \cdot m_{\mathrm{y}} y = m_{\mathrm{y}} Y$, where y was set at full employment. Hence, Walras's law, Say's law and the quantity theory can be logically consistent with each other, without such consistency implying their validity individually or as a set.

than being merely a negatively sloping curve.[15] There can be no meaningful IS–LM or IS–IRT analysis under such a shape of the IS curve. In fact, much of modern short-run macroeconomic theory would be ruled out in such a context.

Therefore, there are many reasons for rejecting the strong form of Say's law – that is, Say's law as an identity – for monetary economies.

18.5 Walras's law, Say's law and the dichotomy between the real and monetary sectors

A *dichotomy* (separation) is said to exist between the real (commodity) and financial (money and bonds) sectors if the demand and supply functions of the former are independent of nominal variables such as the price level, the inflation rate, the nominal interest rate and excess demand in the money and bond markets. In such a case, any shifts in the latter cannot change the values of the real variables under any circumstances.

Say's law alone implies that the demand for commodities always equals their supply, irrespective of the quantities of money, bonds and the price level. Therefore, the "real" system of the economy, which is concerned with the demand and supply of commodities and their relative prices, is independent of financial phenomena in the economy. There is thus a dichotomy between the real and financial sectors of such an economy. Such a dichotomy was also derived in Chapter 3 in the context of the general equilibrium version of the macroeconomic model, under the assumptions that the demand and supply functions were notional and possessed homogeneity of degree zero in all prices, rather than doing so in all prices *and* initial endowments. The present derivation of the dichotomy is related to the earlier one but is from a different perspective, with Say's law embodying within it the homogeneity of degree zero of the demand and supply functions of commodities in all prices. As Chapter 3 has shown, such a dichotomy does not exist in modern financial economies, which possess money and bonds.

Further, as Chapter 3 has argued, the *wealth and the real balance effects* show that a change in the price level changes the real financial wealth (composed of bonds and money balances) of the individual and changes his demand for real money balances and bonds, as well as his demand for commodities. Similar effects would occur if the price level was constant but the money supply was increased in such a way that the wealth of the individual and (the private part of) the economy changed. There is thus interaction between the commodities sector and monetary phenomena through the wealth and real balance effects, so that there does not exist a dichotomy between the real and the monetary sectors. Although these effects were explained in Chapters 3 and 14, we again do so briefly in the following.

18.6 The wealth effect

A change in the real value of wealth induces a *wealth effect* on the demand for commodities by households since this demand by individuals depends on their initial endowments (i.e. wealth). These initial endowments include households' and firms' holdings of money balances and

15 Further, if the economy were only a commodities–money one, then, as argued earlier, Walras's law and Say's law together would imply that the demand for money will also always equal its supply, irrespective of the interest rate and income. Hence, in this case, the LM curve will also span the (r, y) space.

bonds, whose initial values are in nominal or dollar terms, so that a change in the price level of commodities changes the real value of the initial endowments.

In the debates between Keynesians and Keynes's critics in the 1940s and 1950s, the wealth effect is associated with A.C. Pigou, who argued that if aggregate demand was less than full-employment output, prices would fall and increase the real value of wealth. This would in turn increase consumption, thereby increasing aggregate demand and moving the economy to equilibrium at its full-employment level. The impact of the change in the price level on the real value of wealth, and of the latter on consumption, is known as the *Pigou effect*.[16] This effect operates as follows: the economy with deficient demand will continue to generate a price decrease, which will increase the real value of household wealth, which will increase consumption and increase aggregate demand until the demand deficiency is eliminated. While this argument is logical within the context of macroeconomic models, it relies heavily on the *ceteris paribus* condition, which does not normally hold in demand-deficient economies. In fact, in his later writings, Pigou argued that a fall in aggregate demand may not only cause a decline in the price level, it would also bring about simultaneous bankruptcies and deflation, with the consequence that real wealth may fall rather than increase, so that aggregate demand would fall rather than increase.[17] Therefore, the result of the original demand deficiency could more likely be a depression rather than a return to full employment. Hence, while the Pigou effect is an analytical ploy, its practical significance and validity as a device that returns a demand-deficient economy to full employment are doubtful.

At the general level, the wealth effect can occur because of changes in the real values of asset holdings, arising from changes in the nominal values of the assets or of their prices relative to the price level of commodities. The nominal values of the assets and their prices will change if the current or future expected interest rates change. Since changes in the price level can be accompanied by, or themselves induce, changes in interest rates, there can be both direct and indirect changes in wealth connected with changes in the money supply and the price level. Both these changes need to be incorporated in the short-run macroeconomic models. However, stable and predictable relationships between movements of the price level and movements in the prices of bonds, equities and physical assets have not been established. The postulates used for such relationships are at best gross simplifications and reflect a severe deficiency of knowledge on this topic.

18.7 The real balance effect

The real balance effect, associated with Don Patinkin's (1965) contributions from the 1940s to the 1960s, and already discussed in Chapters 3 and 14, is merely one element of the wealth effect and takes account of only those changes in real wealth that arise because of changes in the real value of money balances. This effect was defined in Chapter 3 as the change in the aggregate demand for commodities due to a change in the real value of the money balances held in the economy, with the latter due to a change in the price level or an exogenous change

16 See also the discussion of the Pigou effect in Chapters 3 and 14.

17 Chapter 14 pointed out that Pigou himself considered the effect named after him as a "mere toy" without empirical significance. We have used his argument: a fall in commodity demand will not only produce a price deflation (which increases demand) but, more significantly, it is also likely to cause or increase bankruptcies which will decrease production. The result is more likely to cause a depression than a return to full employment.

in the money stock. Such a change in real balances is part of the change in the individual's wealth due to the change in the price level. The real balance effect is one of the interactions that can occur between the commodity and the money markets and prevents money being neutral in the short run.[18] However, it only applies to outside money (i.e. M0), not inside money (i.e. bank deposits). It is not significant empirically.

18.8 Is Walras's law really a *law*? When might it not hold?

A *law* in economics has been interpreted in this chapter to be a statement that is true as an identity. It must hold under all states of the economy, from the most rudimentary to the most developed states, in recessions and in booms, in normal times and in chaotic conditions. For Walras's law to be a law, we should not be able to adduce any state of the economy, whether common or rare in the real world, which would not obey Walras's law. Walras's law was derived in Section 18.1 from the agents' budgetary constraints. But what would happen if there were also other constraints? Would the particular forms of the additional constraints vitiate Walras's law? The following subsection explores the implications of constraints imposed by the demand in the economy on the actual amounts of output that the firms are able to sell, the actual amounts of labor that households are able to sell and the actual amounts of commodities that households would buy.

18.8.1 Intuition: violation of Walras's law in recessions

The money and bond markets are so efficient in the modern financially developed economies that they adjust continually (every minute) while the financial markets are open, so that they can be taken to be continuously in equilibrium for analytical purposes. That is, we can take $E_m{}^{\mathrm{d}} = E_b{}^{\mathrm{d}} = 0$ *continuously*. Hence, (8) implies that:

$$E_c{}^{\mathrm{d}} + E_n{}^{\mathrm{d}} = 0 \tag{14}$$

so that:

$$E_c{}^{\mathrm{d}} = -E_n{}^{\mathrm{d}} \tag{15}$$

Hence, there must exist positive excess demand for commodities whenever there is excess supply of labor. On a practical note, the evidence of (15) manifests itself in increases in unemployment during recessions, so that, in recessions, $E_n{}^{\mathrm{d}} < 0$. But recessions are also precisely the stage of the business cycle in which firms claim that the demand for their products has fallen and there is not enough demand for them to maintain their employment at the pre-recession levels, so that $E_c{}^{\mathrm{d}} < 0$. That is, the excess supply of labor and the excess supply of commodities occur concurrently in recessions, so that, in recessions, $E_c{}^{\mathrm{d}} + E_n{}^{\mathrm{d}} < 0$. This evidence contradicts (14) and, therefore, Walras's law, thereby casting doubt not only

18 The real balance effect is clearly relevant to the question of whether the economy can be in equilibrium below its full-employment level. This effect was an element in the disputes between the neoclassical economists and the Keynesians on the existence of an equilibrium below full employment, with the neoclassical school arguing that such a state would be one of disequilibrium; prices would decline and changes in aggregate demand and output would occur because of the real balance effect.

on its claim to be a law but also on its validity during recessions for economies with well-developed financial markets. This is a strong indictment of Walras's law and we will pursue its reasoning further in the following analysis.

The following two cases taken from the Keynesian paradigm (Clower, 1965/1969; Leijonhufvud, 1967, 1968, Ch. 2) examine the validity of Walras's law when there exists excess supply in the commodity and labor markets. Both cases assume an economy with efficient financial markets, so that there exists continuous equilibrium in the money and bond markets of the closed economy.

I. Unemployment in the labor market

For Leijonhufvud's arguments, start with the following implication of Walras's law when efficient financial markets ensure continuous equilibrium in the money and bond markets:

$$E^{\mathrm{d}}{}_c = -E^{\mathrm{d}}{}_n \tag{16}$$

Now assume that a shock to the economy produces involuntary unemployment. Since this means that there is excess supply of labor with $E^{\mathrm{s}}{}_n > 0$ (i.e. $E^{\mathrm{d}}{}_n < 0$), Walras's law implies that there must be positive excess demand in the commodity market. That is, firms' response to the rise in unemployment is to increase their production or raise their prices. This is counterfactual, as illustrated by economic analysts' usual interpretation of rising unemployment as an indicator of recession or forthcoming recession, so that their prediction becomes that of a cutback in production by firms.

Let us now follow our intuition and the usual analysis on the plausible and rational micro-economic behavior of households and firms. Since the unemployed workers corresponding to $E^{\mathrm{s}}{}_n$ do not receive any income, they, being rational, reduce their expenditures below those that they would have incurred had they been employed. Given this decrease in consumption expenditures, the commodity market will have an excess supply of commodities, where the supply is determined by what firms would wish to supply if they could sell all that they wanted to sell. In this eventuality, we would have:

$$E^{\mathrm{d}}{}_c < 0 \text{ and } E^{\mathrm{d}}{}_n < 0$$

so that:

$$E^{\mathrm{d}}{}_c + E^{\mathrm{d}}{}_n < 0 \tag{17}$$

Equation (17) contradicts the implication of Walras's law that we must have $E^{\mathrm{d}}{}_c + E^{\mathrm{d}}{}_n \equiv 0$, when the money and bond markets are efficient and adjust continuously to equilibrium. In this scenario, which we have argued above as being quite plausible, the economy displays excess supply of labor without a corresponding excess demand in any market. This violates Walras's law. Hence, the law is not an identity and therefore not a law: it holds when there is equilibrium in all markets but may not hold when there is disequilibrium in the labor market.

What is required in this case for the *reinstatement* of Walras's law? What is required is the assumption that the unemployed workers continue to base their expenditures on the incomes they would have been entitled to receive *if* they had, in fact, been employed. But such posited behavior for households is implausible, irrational and violates the rational

Table 18.1 Walras's Law and the excess supply of labor

	$E^{d}{}_{n}$	$E^{d}{}_{c}$	$E^{d}{}_{m}$	$E^{d}{}_{b}$	$\Sigma_i E^{d}{}_{i}$
Walras's law:	$< 0 \Rightarrow$	$E^{d}{}_{c} = -E^{d}{}_{n}$	0	0	0
		$E^{d}{}_{c} > 0$			
Likely real-world and rational scenario in recessions:	$< 0 \Rightarrow$	$E^{d}{}_{c} < 0$	0	0	< 0
		$E^{d}{}_{c} \neq -E^{d}{}_{n}$			

expectations hypothesis.[19] Alternatively, the reinstatement of Walras's law would require that the firms continue to pay laid-off workers their wages even though they have been laid off. This posited behavior for firms is also implausible, irrational, and violates the rational expectations hypothesis as applied to firms.

We summarize the preceding conclusions in Table 18.1, which has been compiled under the prior assumptions of efficient money and bond markets, so that they possess continuous zero excess demands, and involuntary unemployment.

In Table 18.1, while Walras's law derives $E^{d}{}_{c}$ as an implication from $E^{d}{}_{n}$, the real-world scenario incorporates the *behavioral* relationship between $E^{d}{}_{c}$ and $E^{d}{}_{n}$ imposed by the optimization of its economic agents for the given structure of the economy and their expectations on wage income.

II. Excess supply in the commodity market

For the second case, this time from Patinkin (1965, Ch. 13), start with equilibrium in all markets, so that there is full-employment output. Again, assume that the money and bond markets are efficient, so that they are continuously in equilibrium. Now, let a fall in investment reduce aggregate demand and create excess supply (i.e. $E^{d}{}_{c} < 0$) in the commodity market. Faced with unsold output,[20] firms respond to their unsold output by reducing production and employment until the output of commodities becomes equal to their demand. While this adjustment by firms restores equilibrium in the commodity market through a reduction in their output to match the below-full-employment demand for commodities, it also reduces employment below full employment, so that a state of excess supply emerges in the labor market. Since, given the demand deficiency, there is no reason for firms to increase employment to its initial state of full employment, there continues to be excess supply of labor in the economy. Hence, the economy has excess supply in one market without a corresponding excess demand in any market. This violates Walras's law.

Table 18.2 highlights the arguments of this case on the conflict between Walras's law and the plausible real-world scenario.

19 Rationality requires that economic agents base their decisions on all available information. In the present context, the workers know that they do not receive wage incomes while unemployed, so that they have to decrease their consumption expenditures.

20 Under the perfect market hypothesis for commodity markets, the price will instantly adjust to eliminate excess supply. However, Walras's law does not assume perfect markets, even though the Walrasian general equilibrium model does so. For Walras's law to be an identity, it must hold whether markets are perfect or not.

Table 18.2 Walras's Law and the excess supply of commodities

	$E^{d}{}_{c}$	$E^{d}{}_{n}$	$E^{d}{}_{m}$	$E^{d}{}_{b}$	$\Sigma_i E^{d}{}_{i}$
Walras's law:	$< 0 \Rightarrow$	$= -E^{d}{}_{c}$	0	0	0
		$E^{d}{}_{n} > 0$			
Likely real-world and rational scenario in recessions:	$< 0 \Rightarrow$	$E^{d}{}_{n} < 0$	0	0	< 0
		$E^{d}{}_{n} \neq -E^{d}{}_{c}$			

In Table 18.2, while Walras's law derives $E^{d}{}_{n}$ as an implication from $E^{d}{}_{c}$, the real-world scenario incorporates the behavioral relationship between $E^{d}{}_{n}$ and $E^{d}{}_{c}$ imposed by the optimization of its economic agents for the given structure of the economy and their expectations.

18.8.2 Walras's law under excess demand for commodities

For this section, we again assume that the money and bond markets are efficient and continuously in equilibrium. We now investigate the impact of a shock that produces an excess demand for commodities.

The impact of an excess demand for commodities

Suppose that a positive demand shock, such as to consumption, investment, fiscal deficits and imports, produces an excess demand for commodities. How does the economy usually respond to the increasing demand for its products? The plausible answer is that firms usually respond by adjusting their prices and output, with the output response normally preceding the price response, so that increases in aggregate demand first result in an increase in aggregate output, only later followed by inflation. The increase in production comes about through more intensive uses of both capital and labor, as in the implicit contract theory, as well as through increases in employment. Given the latter effect, the excess demand for commodities produces an excess demand (rising employment) for labor. Hence, the shock that had caused $E^{d}{}_{c} > 0$ led to $E^{d}{}_{n} > 0$, so that:

$$E^{d}{}_{c} + E^{d}{}_{n} > 0 \tag{18}$$

This scenario that $E^{d}{}_{c} > 0$ causes $E^{d}{}_{n} > 0$ is not only plausible, it is commonly observed: central banks and economic analysts use emerging evidence of increasing commodity demand (e.g. in factory orders) to predict a rise in production and a fall in unemployment.

Given the assumption that the money and bond markets are efficient and adjust continuously to equilibrium, (18) implies that $E^{d}{}_{c} + E^{d}{}_{m} + E^{d}{}_{b} + E^{d}{}_{n} > 0$. This contradicts Walras's law, which is that $E^{d}{}_{c} + E^{d}{}_{m} + E^{d}{}_{b} + E^{d}{}_{n} \equiv 0$.

18.8.3 Correction of Walras's law

The preceding analyses imply that the invalidity of Walras's law occurs not only in response to contractionary shocks to aggregate demand, which cause $E^{d}{}_{c} < 0$ and recessions, but also to

expansionary shocks to aggregate demand, which cause $E^{d}{}_{c} > 0$ and booms in the economy. In brief, in the economy, data showing $E^{d}{}_{c} < 0$ is a good basis for a prediction of rising unemployment ($E^{d}{}_{n} < 0$), while data showing $E^{d}{}_{c} > 0$ is a good basis for predicting falling unemployment ($E^{d}{}_{n} > 0$). Such a prediction pattern, commonly used by central banks and economic analysts, runs contrary to the prediction of Walras's law that data showing "high" aggregate demand (i.e. $E^{d}{}_{c} > 0$) would be a good basis for predicting falling unemployment ($E^{d}{}_{n} > 0$), while data showing depressed aggregate demand (i.e. $E^{d}{}_{c} < 0$) would be a good basis for predicting rising unemployment ($E^{d}{}_{n} < 0$).

Since Walras's law does not hold in such disequilibrium states,[21] it is really not a law (identity) but a statement about general equilibrium. Given its lack of validity for disequilibrium, its use in dynamic analysis can lead to misleading conclusions. It can, however, still be used for long-run analysis, since such analysis assumes equilibrium in all markets. Our arguments suggest that the correct statement of Walras's law is:

(i) In general equilibrium, $\sum_{k=1}^{K} E_k^{d} = 0$. However, this statement is trivial since, in general equilibrium, $E_k{}^{d} = 0$ for all k.

(ii) In disequilibrium, $\sum_{k=1}^{K} E_k^{d}$ can be positive (as in booms) or negative (as in recessions).

18.9 Notional demand and supply functions in the classical paradigm

For Walras's law to hold in the above two examples, there must be excess demand for output whenever there is excess supply of labor, and vice versa. Therefore, we need the actual clearance of all markets, or, if some of them are not in equilibrium, we need the hypothetical behavior pattern that *all agents continue to act (and expect) as if all markets cleared even when they do not*. Walras's law requires this assumption, which is unrealistic, to be an identity. As a consequence, the law requires that the demand and supply functions in (1) to (5) be *notional* functions, where a demand or supply function for a good i is said to be notional if economic agents act as if all *other* markets cleared. If this condition does not hold in practice, the operating functions would be *effective* functions, which are based on actual incomes and expenditures and do not assume that all other markets clear. They differ from the notional functions and Walras's law, as specified so far, would not apply to them. Hence, Walras's law would not be an identity for effective functions and strictly does not deserve the designation of a law.

18.10 Re-evaluating Walras's law

18.10.1 Fundamental causes of the failure of Walras's law

The preceding section shows that the assumption of equilibrium in the commodity and labor markets with disequilibrium in the bond and money markets does not lead to a violation of Walras's law. However, as shown in earlier sections, the assumption of continuous

21 Note that with Walras's law as an identity, the arguments against it cannot be rejected by resort to special conditions, such as that of perfect markets. In any case, the assumption of perfect markets is not one of the assumptions of Walras's law.

equilibrium in the bond and money markets with disequilibrium in the commodities and labor markets does lead to such a violation. Why should the economy possess an asymmetry of this type? The reason lies in two positive relationships between commodities and labor. One of these comes from the supply side of the economy, that is, from the production function, which is of the type $y^s = y^s(n)$, so that an increase (decrease) in the production of commodities is accompanied by an increase (decrease) in employment. The other is from the demand side of the economy, for which the demand for commodities can be summarized by $y^d = y^d(n)$, where an increase in employment leads to higher incomes and higher commodity demand. These two relationships individually and together imply a strong positive relationship in practice between the output of commodities and the employment of labor, so that an excess positive demand for output must be accompanied by an excess positive demand for labor, and vice versa. Walras's law fails to take account of these fundamental behavioral relationships, whose implications run counter to the law and lead to its violation.

18.10.2 Irrationality of the behavioral assumptions behind Walras's law

To reiterate, given the starting assumption of continuous bond and money market clearance, Walras's law requires the clearance of both the labor and commodity markets. When they do not clear, it assumes that one of the following two things must occur:

1 Firms continue to produce the full-employment level of output if there is not enough demand for their products, or at least continue to pay workers the full wage even after laying them off.
2 Households behave as *if* the labor market had cleared at full employment, even if it means that unemployed workers spend incomes that they never received.

Both conditions are clearly irrational in the sense of not being profit/utility maximizing under the actual constraints that apply to firms and households, and do not hold empirically.

18.11 Reformulating Walras's law: the Clower and Drèze effective demand and supply functions

18.11.1 Clower effective functions

Clower (1965/1969) argued that in conditions where some markets fail to clear, the relevant demand and supply functions are not notional ones but must take account of disequilibrium in other markets. For example, if some workers are unable to sell their labor supply and are thereby involuntarily unemployed, their constraints on the purchases of commodities, money and bonds must take into account their resulting lack of labor income, so that their actual demands for these goods would be less than their notional demands. That is, their relevant demand functions incorporate the *spillover effects* of disequilibrium in the labor market. However, the relevant supply function of labor is still the notional one since the workers can still buy as much as they like of other goods.

As another example, if firms face deficient demand for commodities, i.e. they cannot sell all the output they wish to produce, they will have an effective demand for labor that is less than its notional demand. However, assuming that the firms do not face disequilibrium in markets other than for commodities, they would still operate with the notional supply

function for commodities as their effective supply function. The demand and supply functions incorporating the impact of disequilibrium in *other* markets are called the *Clower effective demand and supply functions*. In microeconomic analysis, the effective demand/supply of an individual in market i is to be derived from the maximization of his utility function subject to his budget constraint *and* subject to the restrictions perceived by him in all other markets $k, k \neq i$. Similarly, the demand/supply of a firm in market i is derived from its profit maximization subject to its perceived constraints in markets $k, k \neq i$.

Clower (1965/1969) and Leijonhufvud (1967,1968) claimed legitimately that the Clower effective demand and supply functions – and not the notional ones – are the ones applicable to Keynesian analysis, with its emphasis on deviations from full employment, while the notional functions are applicable to classical analysis.[22]

18.11.2 Modification of Walras's law for Clower effective functions

Assuming that a disturbance has led to a fall in employment below the full employment level, we have a constraint in the form $n = n^{\mathrm{d}} \leq n^{\mathrm{s}}$, which modifies equation (1) to the inequality:

$$p_c c^{\mathrm{dh}} + p_c m^{\mathrm{dh}} + p_b b^{\mathrm{d}} + p_e e^{\mathrm{d}} \leq p_c \underline{c}^{\mathrm{s}} + \underline{M}^{\mathrm{h}} + p_b \underline{b}^{\mathrm{s}} + p_e \underline{e}^{\mathrm{s}} + Wn + \pi^{\mathrm{dis}} \tag{19}$$

where $Wn \leq Wn^{\mathrm{s}}$. Inequality (19), in combination with (2) to (5), yields the inequality:

$$E_c{}^{\mathrm{d\#}} + E_m{}^{\mathrm{d\#}} + E_b{}^{\mathrm{d\#}} + E_e{}^{\mathrm{d\#}} + E_n{}^{\mathrm{d}} \leq 0 \tag{20}$$

where the superscript # indicates effective functions,[23] Note that, in the case where $\mathrm{Wn} = \mathrm{Wn}^{\mathrm{s}}$, these functions become notional ones and (20) changes to an equality. Clower argued that the proper statement of Walras's law is (20), that is, the sum of all actual excess demands in the economy is *non-positive*. In this statement of Walras's law, in the context of our assumed involuntary unemployment, the excess demands for goods other than labor are effective ones while the demand for labor is notional. Correspondingly, note that if the perceived constraint by firms was that of deficient demand for commodities – that is, $y^{\mathrm{d}} < y^{\mathrm{s}}$ – the excess demand for commodities would be notional while the excess demands for the other goods would be effective.

A consequence of (20) is that it cannot be used to derive the statement that equilibrium in $K-1$ markets implies equilibrium in the Kth one also, since the $K-1$ markets with demand equal to supply could be those with effective functions while the Kth one could be the one with the disequilibrium, with its disequilibrium being the inequality of its notional demand and supply.

18.11.3 Drèze effective functions and Walras's law

In the context of involuntary unemployment, Drèze (1975) proposed a reinstatement of Walras's law by equating the supply of labor to the effective supply set by the smaller demand for labor. That is, set $n^{\mathrm{s}}{}_{\mathrm{D}} = \overline{n}^{\mathrm{d}}$, where $n^{\mathrm{s}}{}_{\mathrm{D}}$ is the Drèze supply of labor. Its purpose

22 The two are identical if equilibrium exists in all markets.

23 For such functions, a market k can have effective clearance (i.e. $E_k{}^{\mathrm{d\#}} = 0$), with or without notional market clearance (i.e. $E_k{}^{\mathrm{d}} = 0$).

is to impose *all* the constraints that operate in the economy. In the general case, this results in the Clower demand/supply functions for all markets i in which there are no perceived constraints, but with the Drèze demand/supply functions set by the constraints themselves in the constrained k markets. Equations (1) to (5), with demands and supplies redefined in this manner, again imply Walras's law, but in a Drèze effective – and not a notional – sense. Excess demands would also be redefined in a corresponding manner and the sum of all such (Drèze) excess demands would again equal zero.

Note that we can speak of equilibrium in an unconstrained market in the sense of the equality between the Clower demand and supply in that market, and determine its price from such equality. However, we cannot consider the equality of Drèze demand and supply in a constrained market as representing equilibrium in it, or use such equality as a basis for determining the price in it. Further, the Drèze constraint $n^{s}{}_{D} = \overline{n}^{d}$ is strange: it specifies the supply of labor by its demand. There is no intuitive justification for it, as there is for the Clower analysis. Note that both Clower's reformulation of Walras's law as an inequality and Drèze's reinstatement of it in terms of Drèze functions limit the usefulness of this law.

18.12 Implications of the invalidity of Walras's law for monetary policy

What has the preceding discussion on the invalidity of Walras's law in disequilibrium got to do with monetary theory and policy, which is the subject matter of this book?

This chapter has shown that, in disequilibrium with a violation of Walras's law, a shock that produces an aggregate demand deficiency (excess supply) of commodities can be accompanied by an excess supply (unemployment) of labor, and often is. The mechanisms that were adduced to restore such an economy to equilibrium have been the real balance and the Pigou effect. These depend on the demand deficiency producing a fall in the price level, which increases the value of real balances and bonds, causing the wealth effect on consumption to increase aggregate demand in the economy. This process will eventually eliminate the demand deficiency. However, even if such mechanisms were to operate as posited, this process would be extremely slow and may take decades to eliminate a serious demand deficiency. Such a delay is an invitation to the central bank to try to reduce its duration to a much shorter period. The central bank can do so through expansionary monetary policies. These can be reductions in the interest rate or increases in the money supply, but usually both. Such responses to demand deficiency are, in fact, embodied in the Taylor-type rules of monetary policy formulation.

The implication of the invalidity of Walras's law in disequilibrium for a realistic monetary policy is that the relevant dynamic analyses of the movements in output and the price level produced by the economy should be based on the effective excess demand functions, not the notional ones. This is, in fact, the actual practice followed by central banks and economic analysts in predicting future movements in output and prices and the need for an active monetary policy. To illustrate, the output gap in Taylor rules is the deviation of effective (not notional) output from its full-employment level.

Conclusions

Walras's law is a core underlying relationship in the specification of macroeconomic models and is used to eliminate the explicit treatment of one of the markets in them. By convention in

both the classical and Keynesian types of macroeconomic models, the market thus rendered implicit in the analysis is usually the bond market, though it could be any of the others.

Say's law, properly interpreted, is an identity between the supply and the demand for commodities. However, it is not valid for a economy with money and bonds, since such an economy allows substitution between financial assets and commodities – and thereby possesses interaction, at least in the disequilibrium states, between the monetary and commodity sectors.

Walras's law and Say's law imply a dichotomy between the real and monetary sectors and the neutrality of money. The dichotomy is definitely not valid in a monetary economy. The validity of the neutrality of money in the short run depends upon the structure of the economy, since it requires wage and price flexibility, continuous market clearance and the absence of the real balance effect in equilibrium. Few economies meet these stringent conditions in the short run.

While Walras's law is an identity in the context of the *notional* demand and supply functions, it becomes merely an inequality with the Clower effective demand and supply functions. The use of Drèze functions reinstates its equality but at the cost of its usefulness. Keynesian analyses are usually based on Clower effective functions.

The Pigou and real balance effects specify, respectively, the impact of changes in financial wealth and real balances on the aggregate demand for commodities. They are, therefore, among the elements interconnecting the commodity and financial markets and played a critical role historically in discussions on the dichotomy between the sectors. These effects clearly exist in the short run but have limited empirical relevance. The Pigou effect becomes even more doubtful if the impact of a price deflation on interest rates and the insolvency of debt-laden firms are taken into consideration.

Summary of critical conclusions

- Walras's law is based on the budget constraints of the various economic agents in the economy and is perhaps the closest we can get to an identity in economics.
- Say's law does not apply in monetary economies.
- Walras's law can be used to derive the excess demand function for bonds, which is an asset in the macroeconomic framework even though its demand and supply functions are often left unspecified in the standard IS–LM analysis.
- The real balance and Pigou effects are important theoretical links between the money and the commodity markets, but their empirical importance in the modern economy is limited.
- Keynesian analyses of deficient demand and involuntary unemployment use effective demand and supply functions, and modify Walras's law to an inequality.
- Clower and Drèze demand and supply functions are more relevant than notional ones in an economy out of general equilibrium.

Review and discussion questions

1. Walras's law is derived from the budget constraints of all the economic agents in the economy. Can Say's law be similarly derived from budget constraints? Use the relevant constraints and specify the additional assumptions needed for this derivation. Assess the validity of these assumptions.

2. What are the implications for monetary policy if both Walras's law and Say's law are imposed on the IS–LM model? Assess the likely validity of these implications. If they do not seem to be valid, which of these two laws should be discarded? Derive the implications for monetary policy of imposing the remaining law on the IS–LM model.
3. Do the modern classical or/and new classical schools effectively reinstate Say's law as one of their component doctrines? Is so, should they state it explicitly? Discuss.
4. Derive the implications of Walras's law and Say's law together for the determinacy of absolute and relative prices in a commodities–money (no bonds or labor) economy. What role does the real balance effect (in the short run and the long run) play in this determination?
5. Outline your understanding of households' and firms' most likely responses to a fall in the aggregate demand for commodities. Does its perceived duration matter? If a downturn in the economy leads to a fall in demand that is perceived to be significant in magnitude and duration, discuss whether Walras's law will continue to hold. If it does not, what happens to the excess demand for commodities and for bonds in the IS–LM model?

   Does Walras's law hold in recessions?
6. In terms of your understanding and beliefs about the functioning of your economy, which of the four markets (commodities, money, bonds and labor) clear on a daily, weekly and monthly basis? Which may not do so, at least within thirty days of a disturbance? Within 6 months? Within a year?

   If some of the markets do so while others do not do so, does this support Leijonhufvud's (1967,1968) and Clower's (1965/1969) contention that Walras's law is not an identity when there is disequilibrium in some markets?
7. For the magnitudes of the relevant variables in any of the past five years in your economy, try to assess the importance of the real balance effect for a 5 percent fall in the price level.
8. What was the dichotomy between the real and the monetary sectors in the traditional classical approach to macroeconomics? How would such a dichotomy arise in a Walrasian general equilibrium system?

   What was the contribution of Patinkin's real balance effect to this debate?
9. Does the dichotomy between the real and the monetary sectors hold in the modern classical approach? Is this dichotomy valid or not for the modern financially developed economy?
10. What were Keynes's arguments in his attacks on Say's law and the traditional classical dichotomy? In retrospect, and especially in the light of the reversion (counter-reformation!) of macroeconomics to the (modern version of the) classical model, evaluate the success of these attacks.
11. Keynes argued that an economy could be in equilibrium with a substantial amount of involuntary unemployment, but other economists disagreed and argued that a state in which an important market does not clear is one of disequilibrium. Explain the notions of equilibrium and disequilibrium involved, Keynes's justification for his position, and his opponents' justification for theirs. Does the existence of the real balance effect or the wealth effect refute Keynes's position?
12. Discuss: "The real balance effect provides a possible dynamic explanation of the adjustments in the economy in going from one equilibrium to another and is an effective answer to the assertion of a Keynesian under-full-employment equilibrium. However, it is not really necessary for the comparative static propositions of the quantity theory or of neoclassical economics."

13. Discuss: "The derivation using Walras's law of the excess demand function for bonds from those of other markets provides insights into its properties and also shows some clearly invalid assumptions usually made for the excess demand and supply functions for the other markets". Illustrate with examples from IS–LM analysis.
14. Robert Clower argued that "either Walras's law is incompatible with Keynesian economics, or Keynes had nothing fundamentally new to add to orthodox economic theory." Is Walras's law incompatible with the different Keynesian and neoKeynesian models as they have evolved?
15. Define the notional, Clower and Drèze demand and supply functions. Prove whether or not Walras's law applies to each of these three types of functions and in what sense it does so.

References

Baumol, W.J. "Say's law." *Journal of Economic Perspectives*, 13, 1999, pp. 195–204.

Clower, R. "The Keynesian counter-revolution: a theoretical appraisal." In F.H. Hahn and F.P.R. Brechling, eds, *The Theory of Interest Rates*. London: Macmillan, 1965. Also in R.W. Clower, ed., *Monetary Theory, Selected Readings*. London: Penguin, 1969.

Drèze, J.H. "Existence of an exchange equilibrium under price rigidities." *International Economic Review*, 16, 1975, pp. 301–20.

Lange, O. "Say's law: a restatement and criticism." *Studies in Mathematical Economics and Econometrics: In memory of Henry Schultz*. Chicago: University of Chicago Press, 1942.

Leijonhufvud, A. "Keynes and the Keynesians." *American Economic Review Papers and Proceedings*, 57, 1967, pp. 401–10.

Leijonhufvud, A. *On Keynesian Economics and the Economics of Keynes*. New York: Oxford University Press, 1968.

Patinkin, D. *Money, Interest and Prices*, 2nd edn. New York: Harper & Row, 1965.

Shaller, D.R. "Working capital finance considerations in national income theory." *American Economic Review*, 73, 1983, pp. 156–65.

Part VI

The rates of interest in the economy

19 The macroeconomic theory of the rate of interest

The rate of interest is one of the endogenous variables in the Keynesian and classical models, so that its analysis is properly conducted as part of a complete version of those models, which were presented in Chapters 13 to 15.

This chapter singles out the competing views on the determination of the rate of interest and focuses on their differences and validity. It also highlights the very important difference between the comparative static and the dynamic determination of the rate of interest.

Key concepts introduced in this chapter

- Fisher equation of the nominal rate of interest
- Stocks versus flows of funds
- Loanable funds theory
- Liquidity preference theory
- Excess demand function for bonds
- Dynamics of interest rate determination
- Neutrality of money and inflation for the real rate of interest

As explained in Chapters 13 to 15, macroeconomic and monetary analysis until recent decades, and usually even now, has assumed only two distinctive financial assets, money and non-monetary financial assets, and allocated the terms "bonds," "credit" and "loanable funds" as synonyms for the latter. Traditional classical (pre-1936) economists had preferred the term "loanable funds," while modern analysis, as in Chapter 13, prefers the term "bonds" to designate all non-monetary assets. This chapter will use these terms interchangeably, as in Chapters 13 to 15, rather than following the distinctive macroeconomic analysis of Chapter 16, which had two non-monetary financial assets (bonds and credit/loans).

The interest rate is the return on bonds (all non-monetary financial assets), but we have not so far in this book specified explicitly the demand and supply, or the excess demand, functions for bonds. There are two ways of doing so. One is to derive the excess demand for bonds, using Walras's law, from the demand and supply of the other three goods (commodities, money and labor) in the macroeconomic model. The other method is to derive the bond demand and supply functions directly from the behavior of economic agents. We present both of these procedures in this chapter.

Bonds in this chapter comprise a single homogeneous, non-monetary financial asset. Further, to get around issues raised by maturing bonds and to establish a simple relationship

between the nominal bond price p_b and the nominal interest rate R, the (homogeneous) bond is assumed to be a consol (perpetuity), which promises a constant coupon payment of \$1 in perpetuity. For this consol, $p_b = 1/R$.

This chapter studies the comparative static and dynamic determination of the macroeconomic interest rate in the closed economy. In terms of the heritage of ideas, the theories that deal specifically with the determination of this interest rate are the traditional classical loanable funds theory and the Keynesian liquidity preference theory. The loanable funds theory asserts that the bond market determines the interest rate, whereas the liquidity preference theory asserts that the money market does so. The coverage of these theories and their validity is an important part of this chapter.

Given the common or underlying macroeconomic rate of interest as determined in this chapter, the next chapter will examine the time aspects of the interaction among the various interest rates in the economy. The main focus of that chapter will be on the term structure of interest rates.

Section 19.1 reviews the Fisher relationship between the real and nominal interest rates and the impact of inflation on interest rates. Sections 19.2 to 19.6 examine the implications of Walras's law for the determination of the interest rate and the derivation of the excess demand for bonds. Section 19.7 looks at the modern and the historical versions of the loanable funds theory of the rate of interest. Section 19.8 presents the Keynesian liquidity preference theory of interest. Section 19.9 compares the loanable funds and liquidity preference theories in the comparative statics context and shows that the two yield identical comparative static implications for the rate of interest, while their dynamic analyses give different implications, so that a choice has to be made between them. Section 19.10 discusses the neutrality of money for the real interest rate. Sections 19.12 and 19.13 present empirical findings on the Fisher equation, and on the loanable funds and liquidity preference theories.

19.1 Nominal and real rates of interest

The Fisher equation on the interest rate

As explained in Chapter 2, the Fisher equation is:

$$(1 + r^{e}) = (1 + R)/(1 + \pi^{e}) \tag{1}$$

where R is the nominal interest rate, r is the real interest rate, r^{e} is the expected real interest rate and π^{e} is the expected inflation rate. If there exist both real bonds (i.e. promising a real rate of return r per period) and nominal bonds (i.e. promising a nominal rate of return R per period), the relationship between them in perfect markets would be:

$$(1 + R) = (1 + r)(1 + \pi^{e}) \tag{1'}$$

At low values of r^{e} and π^{e}, $r^{e}\pi^{e} \to 0$, so that the Fisher equation is often simplified to:

$$r^{e} = R - \pi^{e} \tag{1''}$$

Mundell–Tobin effect of expected inflation on the real interest rate

In (1), on the interaction between the real interest rate and the expected inflation rate in the context of an exogenous money supply, Mundell (1963) argued, in the context of the

IS–LM analysis with an exogenous money supply, that an increase in the expected inflation rate would cause a reduction in the demand for real balances, which would lower the real interest rate. This came to be labeled the *Mundell effect*. Tobin (1965) argued that, in a general macroeconomic model with a variable physical stock, a reduction in the demand for real balances due to the Mundell effect will increase the demand for real capital,[1] so that as capital accumulates, its productivity falls, which drags down the real interest rate. This is known as the *Tobin effect*.[2] The combined impact of higher expected inflation on the real interest rate is often called the Mundell–Tobin effect.

Impact of high and persistent money growth on the nominal interest rate

Note the impact of changes in the money supply on the nominal interest rate through its impact on the real interest rate and the expected rate of inflation. An increase in the money supply lowers the real rate (the Mundell effect) in the IS–LM analysis through a rightward shift of the LM curve, but it also causes inflation which, through the formation of expectations, creates expected inflation and, through the Fisher equation, raises the nominal rate. At very low rates of expected inflation, the net effect of money creation is often to lower the nominal (and real) interest rates, at least for some time until expected inflation catches up to actual inflation. Over time, high and persistent rates of inflation become expected rates and are invariably accompanied by high nominal rates.

19.2 Application of Walras's law in the IS–LM models: the excess demand for bonds

19.2.1 Walras's law

The derivation of Walras's law was presented in Chapter 18. For the compact four-good (commodities, money, bonds and labor) macroeconomic model, Walras's law is the identity that *the sum of the nominal excess demands for all goods in the closed economy must be zero*. That is,

$$E_c{}^{\mathrm{d}} + E_m{}^{\mathrm{d}} + E_b{}^{\mathrm{d}} + E_n{}^{\mathrm{d}} \equiv 0 \tag{2}$$

where $E_k{}^{\mathrm{d}}$ is the excess *nominal* demand for the kth good, $k = c, m, b, n$. c is consumption, m is real balances, b is bonds and n is labor. Spelled out, (2) is the identity that:

$$P(c^{\mathrm{d}} - c^{\mathrm{s}}) + P \cdot (m^{\mathrm{d}} - m^{\mathrm{s}}) + p_b(b^{\mathrm{d}} - b^{\mathrm{s}}) + W(n^{\mathrm{d}} - n^{\mathrm{s}}) \equiv 0 \tag{3}$$

where the superscripts d and s stand for demand and supply respectively, and:

P = price of commodities (the price level)
p_b = price of bonds
W = nominal wage rate (rental price per period of labor)
c = quantity of commodities

1 With labor, money balances and capital in the production function, fairly standard assumptions imply that a decrease in money balances would increase the demand for both labor and capital.

2 The empirical significance of the Tobin effect is probably negligible in a short-term, as well as a long-term, context. It would require the variability of physical capital in short-run models.

m = real money balances ($= M/P$)
M = nominal money balances
b = quantity of bonds
n = number of workers (labor).

Implications of Walras's law for a specific market

Equation (3) implies that:

$$E_K{}^{\mathrm{d}} \equiv -\sum_{k=1}^{K-1} E_k^{\mathrm{d}} \tag{4}$$

where:

$E_k{}^{\mathrm{d}}$ = excess nominal demand in the kth market
$E_K{}^{\mathrm{d}}$ = excess nominal demand in the Kth market.

Equation (4) allows the excess nominal demand in the Kth market to be derived from the excess nominal demands in the other $(K-1)$ markets. Note that we can arbitrarily designate the market for any specific good as the Kth market. (4) implies that:

$$\text{If } \sum_{k=1}^{K-1} E_k^{\mathrm{d}} = 0, \text{ then } E_K{}^{\mathrm{d}} = 0 \tag{5}$$

which implies the *conditional* statement that *if* there exists equilibrium in $K-1$ markets, *then* there would also be equilibrium in the Kth market. For *comparative static analysis*, (4) allows the analysis to dispense with the explicit treatment of one of the markets in the economy. However, such an omitted market continues to exist and to function but its treatment is pushed into the implicit state. Which market is omitted depends on custom, convenience and the purpose at hand.

Implications of Walras's law for the solution of equilibrium prices

As is clear from (3), the four-good economy has only three prices in it. These are the price level P as the price of commodities, the bond price p_b (or its alter ego, the interest rate) and the nominal wage rate W as the (rental) price of labor. Since Walras's law implies that equilibrium in three markets also ensures equilibrium in the fourth one, we can find the three equilibrium prices by solving the equilibrium conditions for any of the three markets. The calculated equilibrium prices will be invariant to the selection of the three markets for the model being solved.

Walras's law and different groupings of markets in monetary and macroeconomics

The preceding analysis implies that, without changing the equilibrium solution of the three prices, the explicit statement of our macroeconomic model can include only:

(I) money, bond and labor markets;
(II) commodity, money and labor markets;

(III) commodity, money and bond markets;
(IV) commodity, labor and bond markets.[3]

The selection among these choices depends on custom, convenience and the purpose at hand.

The traditional classical (pre-Keynes) economists chose grouping I. They specified the quantity theory for the money market, the loanable funds theory for the bond market and the labor market for the determination of employment. They did not specify a theory for the aggregate demand and supply of commodities and did not explicitly include the commodity market in their analyses.[4] By comparison, following Keynes, the Keynesian school chooses grouping II and specifies the commodity market, the money market and the labor market, but leaves out an explicit analysis of the bond market. The neoclassical and modern classical schools also follow this pattern. However, as we have shown above, as long as the structural equations of the macroeconomic model are the same, Walras's law ensures that the equilibrium values of the three endogenous prices, as well as other real variables such as prices, output, investment, consumption, etc., will be identical among the models. Hence, for comparative static analysis (which solves or compares only the general equilibrium values), the selection of the three markets for explicit analysis is immaterial for the representation of the economy.

An assumption commonly made in modern classical economics is that the labor market is continuously in equilibrium. In this case, Walras's law implies that

$$E_c{}^{\mathrm{d}} + E_m{}^{\mathrm{d}} + E_b{}^{\mathrm{d}} = 0 \tag{6}$$

However, empirically, the underlying assumption of (6) that the labor market continuously clears – while the commodity, money and bond markets may not do so – is highly questionable.

19.3 Derivation of the general excess demand function for bonds

Chapter 13 specified for the open economy the general form of the demand function for commodities by the IS equation as:

$$y^{\mathrm{d}} = y^{\mathrm{d}}(r, P; \lambda) \tag{7}$$

where λ represents the fiscal policy variable, especially the fiscal deficit. The analysis in Chapter 14 of the labor market for the classical paradigm without uncertainty specifies the equilibrium condition for the labor market as:

$$n^{\mathrm{d}}(w) = n^{\mathrm{s}}(w) \tag{8}$$

which determines the equilibrium real wage w, which when substituted in the labor demand function determines the full-employment level n as n^{f}. Substituting the latter in the production function:

$$y = y(n) \tag{9}$$

3 Note that it is rare to find macroeconomic analysis based on groupings III and IV.
4 This is not surprising since the analysis did not include the concept of the consumption function and the multiplier.

determines the supply of commodities y^s as y^f. y^f depends on the supply of labor and technology but is independent of the other variables of the model, so that the classical paradigm's supply function in the absence of uncertainty[5] is:

$$y^s = y^f \tag{10}$$

However, the Keynesian paradigm assumes commodity market imperfections and specifies the commodity supply function by a price/quantity adjustment function, two versions of which are the Phillips curve and the new Keynesian Phillips curve (see Chapter 15). For a comparative statics model, specify the general form of the Keynesian output supply function as:

$$y^s = y(P) \tag{11}$$

Therefore, given the commodity demand function (7) and irrespective of whether we use the commodity supply function of the classical or the Keynesian paradigm, the excess demand ${e_c}^d (= y^d - y^s)$ function for commodities has the general form:

$${E_c}^d = P \cdot {e_c}^d(r, P; \lambda) \tag{12}$$

where ${e_c}^d$ is the commodity excess demand in real terms and ${E_c}^d$ is its nominal value. λ represents the fiscal policy variables.

From Chapter 13, the excess demand function for real balances ${e_m}^d$ is of the general form:

$${E_m}^d = P \cdot {e_m}^d(y, R, P; \theta \cdot \mathrm{M0}, FW_0) \tag{13}$$

where θ.M0 is the money supply, M0 is the monetary base and θ is the multiplier from the monetary base to the money supply. FW_0 is the initial amount of financial wealth.[6]

The analysis of the labor market in Chapter 14 implies the excess demand functions for labor ${e_n}^d$ as:

$${E_n}^d = W \cdot {e_n}^d(w) \tag{14}$$

Hence, by Walras's law as stated in equation (4), the excess real demand function ${e_b}^d$ for bonds is:

$$p_b \cdot {e_b}^d = -P.{e_c}^d(r, P; \lambda) - P \cdot {e_m}^d(y, R, P; \theta \cdot \mathrm{M0}, FW_0) - W.{e_n}^d(w) \tag{15}$$

so that the general form of the excess nominal demand (${E_b}^d$) for bonds is:

$${E_b}^d = {E_b}^d(R, P, w; \lambda, \theta.\mathrm{M0}, FW_0, \pi^e) \tag{16}$$

5 The impact of uncertainty on y^s generates a short-run aggregate supply curve, as shown in Chapter 14.

6 *FW* is generally not specified as an argument of the money demand function. However, omitting it there implies that the excess demand for bonds would not depend on financial wealth, which would be patently invalid and analytically undesirable.

which omits y since y equals $w.n$ in this model. π^e appears as an argument because of the Fisher equation. Alternatively, w can be omitted and replaced by y. Doing so would mean writing the excess nominal demand function for bonds as:

$$E_b{}^d = E_b{}^d(R, P, y; \lambda, \theta.\mathrm{M0}, FW_0, \pi^e) \tag{17}$$

19.4 Intuition: the demand and supply of bonds and interest rate determination

Since R is the nominal return on bonds, its equilibrium value is determined by:

$$b^d(R, \ldots) = b^s(R, \ldots) \tag{18}$$

where b is the number of (homogeneous) bonds/consols. We have assumed that the demand and supply of bonds depend on the nominal interest rate, among other variables. Both the demand for and supply of bonds have a flow and a stock dimension.

Flows versus stocks

In terms of *flows over a specified period of time*, the demand for bonds corresponds to the amount of (loanable) funds flowing or coming onto the market for lending at the various rates of interest. Similarly, the supply of bonds corresponds to the demand for (loanable) funds from those wanting to borrow funds during the period. However, the flow of funds that becomes available for loans over the current period is only a small fraction of the total amount of credit outstanding in the economy. This total amount is like a reservoir and is the *stock* of loanable funds. The stock of loanable funds supplied at any point in time consists of all outstanding loans plus the net additional flow supply of loanable funds, specified for each rate of interest. In stock terms, the demand for credit is similarly the total amount already borrowed plus the net additional amounts that the borrowers wish to borrow at each rate of interest. In modern economies, a major part of this demand often comes from the existing public debt.

In markets with long-term contracts, some of the borrowers and lenders are already committed to loans made at rates prevailing in the past. In such a case, the proper market for determining the current rate of interest is that in terms of flows: the flow market is the actual operating market for bonds in any given period, with borrowers (sellers of bonds) entering it to borrow and lenders (buyers of bonds) entering it to lend funds. However, note that the pre-existing stock of bonds does exert a strong background influence on the flow demands and supplies since parts of this stock of bonds will be expected by borrowers and lenders to mature sooner or later and, over time, become flows available for renegotiation.

The *flow supply of funds* can be interpreted as that part of the stock that has come up for renegotiation plus the additions being made currently. The net *new* supply of funds to the credit market in any period t comes from two sources:

(i) Current (private) saving in the economy.
(ii) Excess supply of money made available for loans, with the excess supply resulting from changes in the public's desired balances or in the supply of money. The supply of money depends on the monetary base and the inside money created by financial intermediaries.

The overall supply of funds in period t is the net new supply from the above two sources plus:

(iii) Funds becoming available from loans that have matured in period t.

The *flow demand* for loans is from net new borrowers and those who wish to renew existing loans. The net *new* demand for loans comes from:

(iv) Current investment in the economy.
(v) Bond-financed government deficits.[7]

The flow demand for loans in period t is from (iv) and (v), plus:

(v) Demand for credit from those whose loans have matured.

Assuming (iii) and (vi) to be equal, the loanable funds theory in flow terms specifies the real demand (f^{d}) and supply (f^{s}) functions of loanable funds as:

$$f^{\mathrm{s}} = s(r, \ldots) + (\theta.\mathrm{M0}^{\mathrm{s}}/P - m^{\mathrm{d}}(R, \ldots)) \tag{19}$$

$$f^{\mathrm{d}} = i(r, \ldots) + (g - t) \tag{20}$$

where:

f^{s} = real flow supply of loanable funds (demand for bonds)
f^{d} = real flow demand for loanable funds (supply of bonds)
s = real saving
i = real investment
g = real government expenditures
t = real government revenues
$\mathrm{M0}^{\mathrm{s}}$ = supply of the nominal monetary base
θ = monetary base to money supply multiplier ($= \partial M^{\mathrm{s}}/\partial \mathrm{M0}$)
m^{d} = demand for real balances
P = price level.

We have assumed that the government deficit ($g - t$) is wholly bond-financed and that r and R are related by the Fisher equation. In partial equilibrium analysis, the equilibrium value of the market rate of interest is determined by:

$$s(r, \ldots) + (\theta.\mathrm{M0}^{\mathrm{s}}/P - m^{\mathrm{d}}(R, \ldots)) = i(r, \ldots) + (g - t) \tag{21}$$

Note that the left side of (21) represents the demand for bonds and the right side represents the supply of bonds. (21) is the statement that the interest rate is determined by the equilibrium in the flow part of the bond market.

Long-run determination of the interest rate

Equation (21) specifies the determination of the short-run interest rate and shows that, although the interest rate is determined by the excess demand for loanable funds, it is

7 This category would be negative for a budget surplus.

not independent of the excess demand for money:[8] excess money demand raises the interest rate and excess money supply lowers it. However, money supply and demand enter the determination of the interest rate *only if* there is disequilibrium in the money market.

In the long run (general equilibrium) the money market would be in equilibrium, so that the excess money demand term on the left side of (21) is zero. Hence, the long-run version of the bond market analysis becomes:

$$s(r, \ldots) = i(r, \ldots) + (g - t) \tag{22}$$

or:

$$s^{\mathrm{n}}(r, \ldots) = i(r, \ldots)$$

where s^{n} is national saving ($= s + (t - g)$). In the context of the closed economy, (22) is also the statement of equilibrium in the commodity sector of the economy.

19.5 Intuition: dynamic determination of the interest rate

We first illustrate the nature of our further arguments by starting with an illustration from the commodity markets. Suppose that equilibrium does not hold in a particular market, say for peanuts. Then the excess demand for peanuts would be a function of the peanut price and of the prices of its substitutes and complements, and the price of peanuts will change in response to the excess demand for peanuts. Further, in general, the larger the excess demand, the faster will be the price change. This adjustment in the price of peanuts is, however, not *directly* influenced by the existence and extent of disequilibrium in the market for other products,[9] even though all markets are related by Walras's law. If there is disequilibrium in a market for a close substitute for peanuts, say potato chips, this disequilibrium will flow into the demand for peanuts, creating disequilibrium in the market for peanuts and influencing their price. But this change in the price of peanuts is not a direct affect of the disequilibrium in the market for chips on the price of peanuts but rather an indirect effect occurring from the spillover, because of substitution, into the demand for peanuts, and depends upon the small or large amount of that spillover. That is, the price of a good responds in a dynamic context directly to the excess demand for that good, with the state of excess demands for other goods being either irrelevant or indirectly relevant in so far as they first affect the excess demand for the good in question.

The rate of interest is a price. But what is the good with which it should be identified in a dynamic context? One approach to this is the liquidity preference concept, explained in detail later, which would identify the interest rate with the good "money," thereby making dynamic changes in the interest rate R a function of the excess demand for money ${E_{mt}}^{\mathrm{d}}$, so that $\partial R_t / \partial t = f({E_{mt}}^{\mathrm{d}})$. The alternative approach is that of loanable funds, which would define

8 Note that the labor market does not appear explicitly in the loanable funds equations (15) to (21), but the determination of incomes in the labor market is clearly a determinant of saving in the economy, so that the labor market is implicitly included in the determination of the loanable funds interest rate.

9 That is, the price of peanuts is not directly a function of the excess demands for other commodities, including, say, almonds, apples or chairs; though the excess demand for peanuts could be indirectly made, through appropriate substitutions, a function of the excess demands for the other commodities.

the relevant good as bonds, thereby making the dynamic changes in the interest rate a function of the excess demand for bonds $E_{bt}{}^{d}$, so that $\partial R_t/\partial t = f(E_{bt}{}^{d})$. These two approaches yield different rates of change in the interest rate.

At a practical level, the dispute between the traditional classical and the Keynesian theories of the rate of interest comes down to which approach will do better empirically in a *dynamic* context. However, there is no generally accepted empirical evidence on this issue, so we should not ignore our intuition on it. Our intuition on the *operational* markets in the economy has already been specified in Chapter 18 and is along the following lines.

In a monetary economy, commodities are always bought and sold at a price against the specific good that is labeled money. There are, therefore, operational markets for commodities (or an operational market for "the commodity" in a single commodity model), so that the equilibrating variables are commodity prices (and the average price level) in commodity markets. Loans are made, again always in a specific good that is money, and the interest rate is agreed between borrowers (sellers of bonds) and lenders (buyers of bonds). It is then the "price" of loans in the bond market. There is, therefore, an operational market for loanable funds (bonds), with the interest rate as the equilibrating variable.

Note that there is no real-world market where money is always bought and sold against *one* specific good. If individuals want to run down their money balances, they can do so either by buying goods in the commodities markets or by making loans in the credit market, with different individuals making this choice in different proportions. These arguments lead to our hypothesis: *the economists' market for money balances is an analytical construct without an operational real world counterpart.*[10] The hypothetical market for money arises only because of Walras's law, a comparative statics concept, and is, as it were, a composite reflection of the other markets. It can be used only for comparative static analysis but not for dynamic analysis to determine the price level or the interest rate in a dynamic disequilibrium context. To conclude, in a dynamic context, the proper analysis of the interest rate is through bond market analysis, as in the loanable funds theory, and not through money market analysis, as in the liquidity preference theory.

The above arguments are buttressed by the buffer stock role of money. As shown in Chapter 6 above, the buffer stock concept is based on the notion that the primary decisions the individuals make are when to change the purchases and sales of commodities and bonds. These decisions result in money holdings that are held passively. By contrast, the "active" decision is not when to buy or sell "money."

19.6 The bond market in the IS–LM diagram

This section uses Walras's law to derive the locus of points in the IS–LM diagram at which there will be equilibrium in the bond market. Figure 19.1 shows the standard IS and LM curves and assumes, for simplification, a closed economy with continuous equilibrium in the labor market with full employment. Given this simplifying assumption, the equilibrium

10 The money market is an "image" provided by the merged reflection of two (or several) entities or figures (markets) standing in front of a mirror, with Walras's law acting as the mirror. Looking at the composite reflection only provides a great deal of information – but not necessarily on the separate elements – about each of the figures (markets) facing the mirror, but is itself not independent of the existence and nature of the mirror, or of the figures (markets).

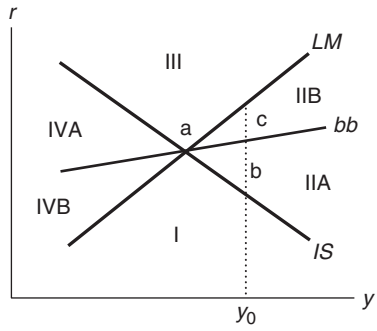


Figure 19.1

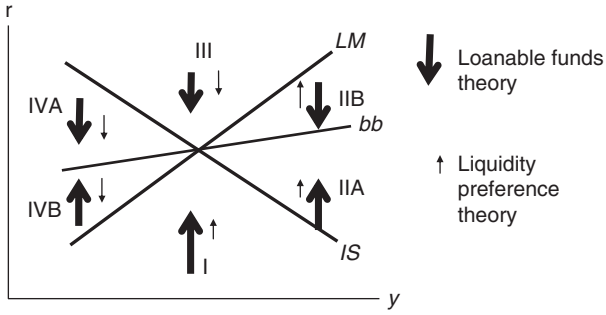


Figure 19.2

curve for the labor market has been omitted from this figure. This figure also assumes an exogenously given expected inflation rate set at zero, so that $r = R$.

The bb curve in Figure 19.1 specifies the combinations of (r, y) which maintain equilibrium in the bond market, so that E_b is zero at all points along it. By Walras's law, the equilibrium in the commodity and money markets shown by the intersection of the IS and LM curves at point a ensures that the bond market will also be in equilibrium at a. Therefore, the bb curve must pass through the intersection of the IS and LM curves, so that all three curves must pass through the common point a.

The IS–LM Figures 19.1 and 19.2 are divided into four quadrants. In quadrant I, there exists excess demand for commodities since, for a given income, interest rates are lower than those specified by the IS curve so that investment is higher than required for equilibrium. That is, $E_c > 0$. There is also excess demand for money, since the interest rate is lower than specified by the LM curve, so that the speculative demand for money is too high compared with that in equilibrium. That is, $E_m > 0$. By Walras's law for our assumed economy, with $E_c > 0$ and $E_m > 0$, we must have $E_b < 0$. That is, there will exist an excess supply of bonds at all points in this quadrant. Therefore, the bb curve (on which, by definition, $E_b = 0$ at every point) cannot pass through quadrant I.

By similar reasoning, it can be shown that in quadrant III, $E_c < 0$ and $E_m < 0$, so that, by Walras's law, we must have $E_b > 0$. Hence, the bb curve also cannot pass through quadrant III.

To summarize, the excess demand functions for quadrants I and III are:

$$\text{I:} \quad E_c > 0, E_m > 0, E_b < 0$$

$$\text{III:} \quad E_c < 0, E_m < 0, E_b > 0$$

Since the bb curve with $E_b = 0$ along it cannot pass through quadrants I and III, it must pass through quadrants II and IV. The latter are further divided into regions IIA and IIB, and IVA and IVB, by the depicted position of the bb curve. To illustrate what happens in regions IIA and IIB, we first examine the point b in Figure 19.1. At this point, the interest rate is too low for equilibrium in the bond market, so that the existing bond prices are too high and there is inadequate demand (excess supply) for bonds at this lower-than-equilibrium interest rate. At point c in the region IIB, the interest rate is too high and bond prices too low for equilibrium in the bond market, so that there would be too much demand (positive excess demand) for bonds at the higher-than-equilibrium interest rate. Similar reasoning can be used to separate regions IVA and IVB. The excess demand functions for regions II and IV are:

$$IIA: \quad E_c < 0, E_m > 0 \text{ and } E_b < 0$$

$$IIB: \quad E_c < 0, E_m > 0 \text{ but } E_b > 0$$

$$IVA: \quad E_c > 0, E_m < 0 \text{ and } E_b > 0$$

$$IVB: \quad E_c > 0, E_m < 0 \text{ but } E_b < 0.$$

Since IIA has $E_b < 0$ while IIB has $E_b > 0$, the separating region between them must have $E_b = 0$. This is the requirement for points on the bb curve, so that the bb curve does differentiate between two distinctive parts of region II. Similarly, IVA and IVB are separated by the locus of points at which $E_b = 0$. Hence, the bb curve passes through regions II and IV and not through I and III. This argument does not establish whether the bb curve will have a positive or a negative slope in the (r, y) space, though we have shown a positively sloping curve. The nature and magnitude of the slope will depend on the coefficients of the IS and LM equations, as can be seen from (21).

We can now examine the impact on the bb curve of changes in the exogenous variables and parameters of the model. We do so explicitly only for the policy variables of the fiscal deficit and the money supply. Start with an increase in the fiscal deficit in Figure 19.1. This would shift the IS curve to the right (not shown). With the LM curve taken to be independent of the fiscal deficit, by Walras's law, the bb curve must shift to pass through the intersection of the new IS and the initial LM curves. Hence, the bb curve will also shift to the right. The intuitive reason has to do with the bond financing of the deficit, which increases the supply of government bonds in the economy. Equilibrium in the bond market requires either higher income or higher interest rates to generate a corresponding increase in the demand for bonds.

But suppose, instead, that the money supply had increased. This would shift the LM curve to the right. Walras's law again implies that the bb curve must shift to pass through the new intersection of the initial IS curve and the new LM one. Hence, the bb curve shifts downwards. The intuitive reason for this is that the increased money supply is traded by the public for bonds, thereby increasing the demand for bonds. This raises bond prices and lowers the interest rate.

The dependence of the bb curve on shifts in both the IS and the LM curves is also apparent from equation (21), which shows that the excess demand for bonds depends on the fiscal deficit and the money supply. The bond market is thus the implicit link between the IS and the LM curves in the IS–LM model.

We derived the bb curve under the assumption of full employment and therefore of constant real output. If the labor market were in disequilibrium, Walras's law would require a modification of our arguments, without, however, a change in the nature of the bb curve. Further, in general equilibrium, the labor market would clear, so that the bb curve must still pass through the intersection of the IS and LM curves.

19.6.1 Diagrammatic analysis of dynamic changes in the rate of interest

The loanable funds theory asserts that the dynamic movement of the interest rate is determined by the excess demand for bonds, while the liquidity preference theory asserts that it is determined by the excess demand for money. Figure 19.2 uses arrows to show the movements in the interest rate implied by the liquidity preference theory, which asserts that $\partial R/\partial({E_m}^{\mathrm{d}}) > 0$, and the loanable funds theory, which asserts that $\partial R/\partial({E_b}^{\mathrm{d}}) < 0$. The movements implied by the former theory are shown by light arrows indicating the direction of movement, while those implied by the latter are shown by the heavy arrows. Both theories predict an increase in the interest rate in quadrant I, and both theories predict a decline in it in quadrant III. Further, in region IIA, both theories indicate a rise in the interest rate, and in region IVA both theories indicate a fall in it. However, note that the implied magnitude of the change in the interest rate could differ between the theories.

The especially interesting regions are IIB and IVB. Region IIB has $E_m > 0$ and $E_b > 0$. Therefore, the liquidity preference theory predicts a rise in the interest rates while the loanable funds theory predicts a fall. In region IVB, $E_m < 0$ and $E_b < 0$, so that the liquidity preference theory predicts a fall in the interest rate and the loanable funds theory predicts a rise. Consequently, the dispute between these theories is not trivial even in terms of the sign of the movements in the interest rates in the economy.

The dispute is of importance even in quadrants I and III and regions IIA and IVA, where the two theories predict a similar direction of change in the interest rate, but the dynamic speed of movements in the rate will be sensitive to the actual relationship and is likely to differ. Hence, a choice has to be exercised between the two theories for both qualitative and quantitative dynamic analyses, though not for the general equilibrium case.

19.7 Classical heritage: the loanable funds theory of the rate of interest

The traditional classical economists (prior to Keynes) had generally favored the specification of the overall equilibrium in terms of the bond, money and labor markets, with the labor market determining employment and, through the production function, output; the bond market determining the interest rate, and the market for money determining the price level. Its theory of the determination of the interest rate (or its inverse, the price of bonds for bonds specified as consols which have fixed coupon payments payable perpetually) was known as the *loanable funds theory*. It asserted that the interest rate was determined in the bond market by equilibrium between the demand and supply of "loanable funds," which was its synonym for the current term "bonds." Given the discussion so far

in this chapter, we can distinguish the following three aspects of the loanable funds theory:

1 Partial equilibrium (short-run) determination of the interest rate.
2 General equilibrium (long-run) determination of the interest rate.
3 Dynamic movement of the interest rate.

The traditional classical economists did not have a distinction (until the advent of Fisher's equation) between the real and nominal interest rate, so that they normally referred to the market interest rate R as the determinant of investment and saving. Following their pattern of analysis, we will specify the investment function in the following as $i(R)$, rather than our usual $i(r)$.[11] They also did not have a government sector with outstanding bond-financed budget deficits.[12] Further, the role of financial intermediaries and the monetary base in the money creation process were not fully understood.[13]

Given these simplifications, the demand and supply functions of the loanable funds theory were:

$$Pf^{s} = Ps(R, y) + [M^{s} - M^{d}]$$

$$Pf^{d} = Pi(R)$$

Therefore, *the short-run* (i.e. partial equilibrium) *determination of the interest rate* according to the loanable funds theory was specified by:

$$s(R, y) + (1/P)[M^{s} - M^{d}] = i(R) \tag{23}$$

so that:

$$R = \varphi(P, y; (M - M^{d})) \tag{24}$$

which allows both the commodity market *and the money market* shifts to change the interest rate.

The *long-run version of the loanable funds theory* assumed general equilibrium in the economy. Therefore, for this version, $M - M^{d} = 0$, so that the long-run loanable funds theory became:

$$i(R) = s(R, y) \tag{25}$$

11 This would also be correct under the Fisher equation if the expected inflation rate was zero.

12 The loanable funds theory was formulated in a period when the size of the government was relatively small. In any case, the formal bond market was not sufficiently developed for governments to sell bonds in them. Often, any borrowing by the government was through private arrangements with individual banks and private financiers.

13 The traditional classical theories were also formulated for an era in which the formal financial sector was relatively insignificant. Until the Second World War, the market for credit was dominated by firms ("ultimate borrowers") raising funds for investment and savers ("ultimate lenders") lending out of savings, so that it was common for the role of financial intermediaries to be ignored or, in any case, not properly integrated into the theory of the rate of interest.

Further, long-run equilibrium in the commodity and labor markets ensures that output will be at the full-employment level (y^f), so that (25) becomes:

$$i(R) = s(R, y^{\mathrm{f}}) \tag{26}$$

That is, the long-run interest rate is determined by the equality of investment and saving at full-employment output, so that its main determinants are the propensity to save, the production capacity of the economy, and investment. In particular, this interest rate is not altered by shifts in the demand or supply of money.

For the *dynamic movement of the interest rate* when there is disequilibrium in the economy, the loanable funds theory asserted that the interest rate is determined by the excess demand or supply of loanable funds: it falls (while the bond price rises) if there is an excess demand for loanable funds, and rises (while the bond price falls) if there is an excess supply of loanable funds. Therefore, this theory's assertion for dynamic adjustments in the interest rate is:

$$R = f(E_b{}^{\mathrm{d}}) \qquad \partial R/\partial(E_b{}^{\mathrm{d}}) < 0 \tag{27}$$

Note that, since changes in the money supply and demand alter the excess demand for bonds, they will also affect the dynamic path of the interest rate.

To conclude on the relevance of the excess money supply in changing the nominal interest rate, only the long-run version of the loanable funds theory asserted its irrelevance for the determination of the interest rate. However, this long-run version, which is the statement that the interest rate is determined by saving at full employment and investment, is the one usually remembered as the statement of the loanable funds theory.

Adapting the loanable funds theory to the modern economy requires the introduction of the government sector, the central bank and the financial sector into the component functions of this theory, as presented earlier in this chapter.

19.7.1 Loanable funds theory in the modern classical approach

In recent years, the modern version of the classical paradigm has reasserted continuous market clearance for the labor markets, as for the other markets, and with its assumption of rational expectations has further asserted the possibility of disequilibrium (due to expectational errors) in any market as at best a very transitory state. That is, with the labor and money markets clearing continuously, there would exist full employment in the economy and the excess demand in the money market would be zero. Consequently, for the modern classical school, the theory of interest reverts to the long-run version of the traditional loanable funds theory, with the difference that it is now intended to be not only the long-run theory but also the short-run theory of the rate of interest as far as systematic or anticipated changes in the money supply are concerned.[14] However, such a short-run theory could still diverge from its long-run version because of random influences operating on the economy in the short run. These cannot be anticipated under rational expectations and would cause a divergence of the short-run interest rate from its long-run level.

14 The latter is a departure from traditional classical economics, as a comparison of the doctrines of the modern classical school with the quotation from Hume in the next subsection clearly shows.

The modern classical version of the loanable funds theory, therefore, extends and differs from the traditional classical one in various ways. Among the differences are:

1 The role of financial intermediaries, as discussed earlier.
2 The addition of Fisher's equation connecting the real and nominal interest rates in perfect capital markets.
3 The distinction in the modern version between the anticipated and unanticipated values of the relevant variables, among which are the money supply, the other determinants of aggregate demand and the rate of inflation. Anticipated money supply increases cause anticipated inflation without changing the real interest rate and, therefore, increase the nominal rate by the anticipated rate of inflation, as specified by the Fisher equation. Unanticipated money supply growth lowers the real rate and will lower the market rate of interest.
4 Ricardian equivalence, which makes national saving independent of the (anticipated) fiscal deficit and therefore removes such deficits from the determinants of the demand and supply of loanable funds. In this case, anticipated deficits would not affect the interest rate.
5 In the short run, the traditional classical economists allowed deviations from full employment under the impact of money supply changes and the impact of these changes on saving. The modern classical economists allow such a deviation for only unanticipated money supply changes. Therefore, the short-run deviations of output from its full employment level under the impact of anticipated money supply changes could, in the short run, affect the interest rate under the traditional version of the loanable funds theory but not under its modern version.

Note that outside the confines of long-run general equilibrium analysis, the interest rate is not merely the reward for postponing consumption, it is also the return on lending, which is the act of parting with liquidity, i.e. not holding money. The latter was a major contention of Keynes and is a fundamental part of Keynesian beliefs.

19.7.2 David Hume on the rate of interest

David Hume occupies a special place in the macroeconomic theory of the rate of interest because he specified the main elements of the preceding traditional classical theory at an early stage in its development. He expressed some of the basic elements of the traditional classical short- and long-run theories in his essay *On Interest*, published in 1752. He stated its long-run version as:

> For, suppose that, by miracle, every man in Great Britain should have five pounds slipped into his pocket in one night; this would much more than double the whole money that is at present in the kingdom; yet there would not next day, nor for some time, be any more lenders, nor any variation in the interest. And were there nothing but landlords and peasants in the state, this money, however abundant, could never gather into sums, and would only serve to increase the prices of everything, without any further consequence. The prodigal landlord dissipates it as fast as he receives it; and the beggarly peasant has no means, nor view, nor ambition of obtaining above a bare livelihood. The overplus of borrowers above that of lenders continuing still the same, there will follow no reduction of interest. That depends upon another

> principle; and must proceed from an increase of industry and frugality of arts and commerce.
>
> The greater or less quantity of it [money] in a state has no influence on the interest. But it is evident that the greater or less stock of labor and commodities must have a great influence; since we really and in effect borrow these, when we take money upon interest.
>
> (Hume, *Of Interest*, 1752).

Hence, for the long run, Hume asserts that changes in the money supply have no impact on the (long-run) interest rate, which is determined by the real factors of labor supply, productivity of investment, and saving. For the short run, they also have no impact if increases in the money supply are spent wholly on commodities but not saved. Hume next considers the *dynamics* of a change in the money supply in the following statements.

> Another reason of this popular mistake with regard to the cause of low interest, seems to be the instance of some nations, where, after a sudden acquisition of money, or of the precious metals by means of foreign conquest, the interest has fallen. … it is natural to imagine that this new acquisition of money will fall into a few hands, and be gathered into large sums, which seek a secure revenue, either by the purchase of land or by interest. … The increase of lenders above the borrowers sinks the interest, and so much the faster if those who have acquired those large sums find no industry or commerce in the state, and no method of employing their money but by lending it at interest. But after this new mass of gold and silver has been digested, and has circulated through the whole state, affairs will soon return to their former situation. … The whole money may still be in the state, and make itself felt by the increase of prices; but not being now collected into any large masses or stocks, the disproportion between the borrowers and lenders is the same as formerly, and consequently the high interest returns.
>
> (Hume, *Of Interest*, 1752).

The salient points of Hume's conclusions can be summarized in modern terminology as follows.

1. The long-run real equilibrium interest rate is determined by saving at full-employment output and productivity of investment.
2. The long-run interest rate is invariant to changes in the money supply. A long-run decline of the rate of interest is brought about by a decline in the productivity of investment and not by increases in the money supply.
3. However, in the short run, the real interest rate is lowered (and output increased) by those increases in the money supply that increase the supply of loanable funds in the economy, as occurs in the indirect transmission mechanism, but not by those that do not, as occurs in the direct transmission mechanism. The structure of the economy and the mode of introducing additional money balances into the economy determine which mechanism is the relevant one, and how long the reduction of the interest rate will last.

Note that the only thing that seems to be missing from Hume's arguments is the distinction between the real and the nominal interest rates: Hume did not have Irving Fisher's equation, proposed early in the twentieth century, relating the nominal interest rate to the expected rate of inflation.

19.8 Keynesian heritage: the liquidity preference theory of the interest rate

Keynes's *General Theory* (1936) challenged the loanable funds theory on the grounds that the interest rate was not the reward for saving but was rather an inducement to part with liquidity. He summarized his views in the statement:

> [Once the individual has made his decision on consumption versus saving out of his income], there is a further decision which awaits him, namely, in what form he will hold the command over future consumption which he has reserved, whether out of his current income or from previous savings. Does he want to hold it in the form of immediate, liquid command (i.e. in money or its equivalent)? Or is he prepared to part with immediate command for a specified or indefinite period. …
>
> It should be obvious that the rate of interest cannot be a return to saving or waiting as such. For if a man hoards his savings in cash, he earns no interest, though he saves just as much as before. On the contrary,…, the rate of interest is the reward for parting with liquidity for a specified period. …
>
> Thus the rate of interest at any time, being the reward for parting with liquidity, is a measure of the unwillingness of those who possess money to part with their liquid control over it. The rate of interest is not the "price" which brings into equilibrium the demand for resources to invest with the readiness to abstain from present consumption. It is the "price" which equilibrates the desire to hold wealth in the form of cash with the available quantity of cash. … If this explanation is correct, the quantity of money is the other factor, which, in conjunction with liquidity preference, determines the actual rate of interest in given circumstances.
>
> (Keynes, 1936, pp. 166–8).

First, consider Keynes's argument in terms of its general notion that the interest rate is the reward for parting with liquidity. This is definitely true in a world with uncertainty. Savers have a choice as to the form in which to hold their savings. They may hold these in a monetary form or lend it. If the level of the interest rate determines their division of savings into money balances versus loans, the interest rate can be called the reward for parting with liquidity in the process of lending. However, if the interest rate also influences the level of savings, then it may also be called a reward for postponing consumption. Both cases apply in the real world.[15]

Now consider Keynes's argument formally in terms of the equilibrium relationship of the monetary sector. As shown in Chapter 2 above, Keynes's money market equilibrium relationship for an exogenously given money supply M is:

$$M = kPy + L(R) \tag{28}$$

Equation (28) determines R *if* it is assumed that P and y are exogenously given. This is not true of the Keynesian model and is not true for Keynes's own ideas in general. In his theory, output, interest and prices were determined simultaneously so that R is not determined merely by (28): it is also influenced by the saving and investment decisions of the expenditure sector

15 However, several empirical studies show the impact of interest rates on saving to be insignificant or of little importance.

as well as by the labor market structure. Hence, the interest rate is not merely the reward for parting with liquidity, even though that may seem to be the most proximate or closely related cause.

Dynamics of the liquidity preference theory

According to Keynes's liquidity preference theory, the dynamic movements of the interest rate are determined by the excess demand for money. Hence, it was asserted that:

$$R = f(E_m{}^{\mathrm{d}}) \qquad \partial R/\partial(E_m{}^{\mathrm{d}}) > 0 \tag{29}$$

so that:

$$\partial R/\partial(M^{\mathrm{d}} - M^{\mathrm{s}}) > 0 \tag{30}$$

The argument behind this assertion runs as follows. According to Keynes an excess demand for money by individuals would make them sell bonds in order to obtain the extra money balances they want. These bond sales will lower bond prices, which will raise the interest rate.

19.9 Comparing the liquidity preference and the loanable funds theories of interest

A lengthy controversy flared up in the 1950s and early 1960s as to whether the traditional classical loanable funds and the Keynesian liquidity preferences theories were identical or different. Given our analysis so far, this question can be answered for the separate categories of general equilibrium and dynamic analyses.

General equilibrium analysis

Our earlier analysis implies that, given Walras's law, it is immaterial whether the *general equilibrium solution* of the macroeconomic model is obtained from (i) the traditional classical set consisting of the money, bond and labor markets, or (ii) the Keynesian set consisting of the commodities, money and labor markets. Each set would give the *same* general equilibrium values of all the endogenous variables, even though the two sets, *prima facie*, would seem to be quite different.

Note that the current practice in macroeconomic modeling is to specify the complete model in terms of the commodity, money and labor markets. The selection of the bond market for omission is partly due to the tradition set by Keynes in *The General Theory*, reinforced by Hicks in his interpretation of Keynes in the form of the IS–LM model, and partly because most countries do not publish adequate and reliable data on the amounts of bonds in the aggregate and for many types of bonds (and loans) in the economy. By comparison, the data on output, money and employment – and their related variables – is usually made available in great detail and with more or less of an attempt at consistency over time.

Dynamic analysis

For the dynamics of interest rate movements, the selection of the sector in which the rate of interest is determined is highly relevant. This has already been shown above in various places. The following pulls it together in one place.

The dynamic version of the liquidity preference theory is that the changes in the rate of interest are determined by the excess demand for money, with a positive relationship. That is:

$$\begin{aligned} \mathrm{d}R/\mathrm{d}t &= \Psi(E_{mt}) \qquad \Psi' > 0 \\ &= \Psi(M_t^{\mathrm{d}} - M_t^{\mathrm{s}}) \end{aligned} \tag{31}$$

The dynamic version of the loanable funds theory is that changes in the rate of interest are determined by the excess demand for bonds, with an inverse relationship:

$$\begin{aligned} \mathrm{d}R/\mathrm{d}t &= \phi(E_{bt}) \qquad \phi' < 0 \\ &= \phi(P\{i(R) - s(R)\} - \{M^{\mathrm{d}}{}_t - M^{\mathrm{s}}{}_t\}) \end{aligned} \tag{32}$$

Equations (31) and (32) generate different time paths for the interest rate. To illustrate an extreme scenario, if the money market is in equilibrium but the bond market is not, the liquidity preference theory (31) would have $E_{mt} = 0$ and imply that the interest rates will not change, but the loanable funds theory (32) would have a non-zero excess demand for commodities and bonds and would, therefore, imply changes in the interest rate.

19.10 Neutrality versus non-neutrality of the money supply for the real rate of interest

Neutrality of money under an exogenous money supply

The real rate of interest is a real variable. As such, the modern classical analysis (Chapter 14) implies that in the long-run general equilibrium (without errors in price or inflationary expectations) the real interest rate will be invariant to changes in the (nominal) money supply, since such changes in the money supply produce proportionate changes in the price level without changing the real money supply. Hence, changes in the money supply do not change the long-run real interest rate in the modern classical analysis.

However, for the short run, the modern classical school allows errors in expectations to affect the real interest rate. This effect can be illustrated by the adaptation of the Friedman–Lucas output supply function to interest rate determination, as in:

$$r_t = r^* + \gamma(M_t - EM_t) \qquad \gamma < 0 \tag{33}$$

where r^* is the long-run value of the real rate of interest. Hence, in the modern classical models, unanticipated increases in the money supply are not neutral in the short run; they lower the real interest rate. But anticipated increases in the money supply do not change the real interest rate and are neutral.

The new Keynesian models incorporate several elements, such as imperfect information and sticky prices, which lead to the non-neutrality of money with respect to output and employment. They also do so for the real interest rate.

Neutrality of monetary policy under a Taylor interest rate rule

Chapter 13 presented the general form of the Taylor interest rate rule as:

$$r^{\mathrm{T}}{}_t = r^{\mathrm{LR}} + \alpha(y_t - y^{\mathrm{f}}) + \beta(\pi_t - \pi^{\mathrm{T}}) \qquad \alpha, \beta > 0 \tag{34}$$

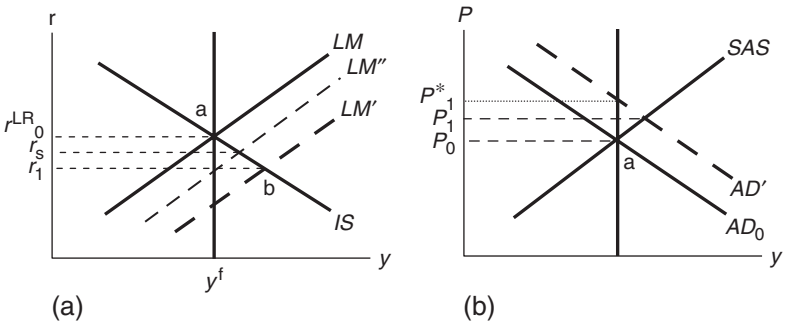


Figure 19.3

where r^T is the real interest rate target, y is real output, y^f is full-employment output, π is the actual inflation rate, π^T is the inflation rate desired by the central bank, and the subscript t refers to period t. π^T is called the *target inflation rate*. As shown in Chapter 14, in the long run, $y_t = y^f$ and $\pi_t = \pi^T$, so that $r = r^{LR}$. Hence, the long-run interest rate will be invariant with respect to monetary policy even if the central bank sets interest rates. However, in the short term, the real rate in the financial markets will depend on the interest rate set by the central bank, so that monetary policy will not be neutral in terms of its effect on the real interest rate on bonds.

Diagrammatic analysis of the role of the money supply in the determination of the rate of interest

Figure 19.3a shows the long-run equilibrium interest rate at $r^{LR}{}_0$. In the long run, an increase in the money supply will shift the LM curve out to LM′, increasing demand to the point b, thereby causing a long-run price increase (from P_0 to $P^*{}_1$ in Figure 19.3b) which is sufficient to return the LM curve in Figure 19.3a back from LM′ to LM and the general equilibrium real interest rate back to $r^{LR}{}_0$. The long-run equilibrium real interest rates and output are therefore invariant to the nominal money supply increase.

But, in the short run, assuming that the money market adjusts instantly while the commodity market is slower to adjust, the economy would initially lower the real interest rate to r_1. If the economy proceeds along a short-run supply curve SAS – either for Keynesian or for modern classical reasons (with errors in relative price expectations) – the monetary expansion will shift the demand curve to AD′ in Figure 19.3b, and lead to a price increase from P_0 to P_1 (rather than to $P^*{}_1$). This will mean, in Figure 19.3a, a shift in the LM curve up from LM′ to only LM′ (rather than back to LM), and yield a short-run rate of interest r_s. Therefore, increases in the money supply can have both immediate (from $r^{LR}{}_0$ to r_1) and short-run effects (from r_s to $r^{LR}{}_0$) on the interest rate in the economy, but not in the long run.

19.11 Determinants of the long-run ("natural") real rate of interest and the non-neutrality of fiscal policy

We now look at the factors that can change the long-run real rate of interest. Figure 19.4 shows the determination of the long-run real interest rate r^{LR}. This is given by the intersection of the

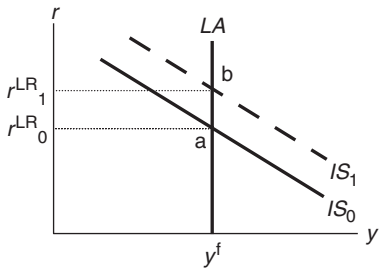


Figure 19.4

IS curve and the long-run aggregate supply curve LAS. Shifts in either of these will change the long-run rate of interest in the economy. Therefore, productivity shifts in the economy as well as its saving and investment behavior will shift the long-run real rate of interest. So will government expenditures and tax rates in the standard IS–LM model, since a deficit in this model shifts the IS curve to the right. This effect is shown in Figure 19.4 with a rightward shift in the IS curve from IS_0 to IS_1 due to a budget deficit, with a consequent increase in the long-run interest rate from $r^{LR}{}_0$ to $r^{LR}{}_1$. That is, higher deficits produce higher real rates of interest, and eliminating them will lower the real interest rate in the long run. The LM curve is irrelevant to the determination of the long-run real rate of interest, so that it has not been drawn in Figure 19.4. We conclude from this figure that, for the closed economy and a given production function, the long-run real interest rate is determined by the equality of investment and national saving, which equals private saving less the fiscal deficit. A decrease in national saving reduces the loanable funds available for investment and raises the real interest rate.

However, the addition of Ricardian equivalence to the IS–LM framework implies, as shown in Chapters 13 and 14, that private saving increases by the amount of the deficit, so that national saving does not change as a result of a deficit. Therefore, budget deficits will not shift the IS curve, so that, in Figure 19.4, the long-run real interest rate will remain at $r^{LR}{}_0$ irrespective of the deficit. Hence, given Ricardian equivalence, the long-run real interest rate will be invariant to fiscal policies. A cautionary note about this conclusion is needed: if Ricardian equivalence does not hold (see Chapter 14 on its doubtful empirical validity), fiscal deficits do alter the long-run real interest rate.

Other determinants of the long-run real rate of interest

The other determinants of the long-run real interest rate include the following.

1 The efficiency of financial intermediation and innovations in it, since these affect the efficiency with which savings are collected and allocated among investment projects.
2 Innovations and the rate of technical change generally in the economy, which affect the growth of output.
3 The openness of the economy to world markets and international capital flows, since such flows can cover a saving–investment gap.

19.12 Empirical evidence: testing the Fisher equation

For tests of the Fisher equation of the nominal interest rate, note that an unanticipated increase in the money supply lowers the short-run real interest rate as the economy moves down the SAS curve, thereby tending to lower the nominal rate, while the resulting inflation, through anticipations, causes an increase in the nominal rate. Further, the Mundell–Tobin effect, discussed earlier in this chapter, implies a negative impact of the expected inflation rate on the real interest rate since the former is the opportunity cost of holding money and reduces the demand for real balances.

A test of the Fisher equation is provided by Crowder and Hoffman (1996). Since an increase in the interest rate due to inflation increases the tax payments on interest receipts, Crowder and Hoffman argued that the empirical estimates of the Fisher equation should have an estimated coefficient of the rate of inflation between about 1.3 and 1.5, rather than unity, as implied by the tax-free Fisher equation, even though many empirical studies had found this coefficient to be less than unity. Since the data on nominal interest rates and inflation tends to be non-stationary, Crowder and Hoffman use the Johansen cointegration technique. Their estimating equation is based on a generalized form of the Fisher equation derived from intertemporal utility maximization subject to a budget constraint, and is stated as:

$$R_t(1-\tau_t) = r_t + E_t\Delta p_{t+1} + 0.5\ \text{var}_t\Delta p_{t+1} - \gamma\ \text{cov}_t(\Delta c_{t+1}, \ldots, \Delta p_{t+1}) \tag{35}$$

where p is the log of the price level, c is the log of consumption, E_t is the expectations operator conditional on information in t, R is the nominal rate and r the real one, τ is the tax rate and γ is the coefficient of relative risk aversion. In this study, the estimated coefficients for the expected rate of inflation were in the range from 1.34 to 1.37 and, when adjusted for the tax rates, in the range 0.97 to 1.01, so that this study supported the Fisher effect.

The Fisher effect incorporated into the theory of the term structure of interest rates allows the estimation of the expected rate of inflation from the yield curve. This derivation will be discussed in the next chapter.

19.13 Testing the liquidity preference and loanable funds theories

Finding an empirical test that would truly distinguish between the liquidity preference and the loanable funds theories poses quite a problem. The strong distinction between these theories only emerges in the hypothetical/analytical long-run general equilibrium state of the economy, which is extremely difficult to test for since the available data never relates to this state of the economy. The tests therefore have to be of the theories' short-run versions. But the short-run versions of both theories imply that excess money demand does affect the interest rate. To illustrate the nature of the problem, the loanable funds theory (LF) asserts that:

$$\text{LF:} \quad R = f(E_b) \tag{36}$$

which, by Walras's law with labor market clearance, becomes:

$$\text{LF:} \quad R = f(-E_c - E_m) \tag{37}$$

The liquidity preference theory (LP) asserts that:

$$\text{LP:} \quad R = f(E_m) \tag{38}$$

Hence, E_m occurs as a determinant of the interest rate in both theories.

Alternatively, using Walras's law to replace E_m by $(-E_c - E_b)$, we have:

$$\text{LF:} \quad R = f(E_b) \tag{39}$$

$$\text{LP:} \quad R = f(-E_c - E_b) \tag{40}$$

In this case, E_b occurs as a determinant of R in both theories, even the liquidity preference theory.

Therefore, the problem in estimations meant to distinguish between the loanable funds and liquidity preference theories arises because both E_b and E_m affect the interest rate in both theories.

Empirical findings

Feldstein and Eckstein (1970) sought to provide an application of the liquidity preference theory, combining it with the Fisher equation for the relationship between the nominal interest rate and the expected inflation rate. The liquidity preference theory implies that the rate of interest depends in the short run upon the excess demand for money. Since the demand for money depends upon income, the excess demand for money was represented through the use of the monetary base and national income among the explanatory variables, with the former capturing the effect of an increasing money supply and having a negative expected coefficient, while the latter captures the effect of increasing money demand and has a positive expected coefficient. The expected rate of inflation was proxied by a distributed lag autoregressive model. Among the results reported by these authors are:

$$R_t = -11.27 - 6.76 \ \ln \ \mathrm{M0}_t + 6.03 \ \ln \ y_t + 0.275\pi_t + \Sigma_j \alpha_j \pi_{t-j} \tag{41}$$

where $j = 1, 2, \ldots, 23$, $\Sigma_j \alpha_j = 3.41$, with $\alpha_1 = 0.289$ and $\alpha_{23} = 0.020$, $R^2 = 0.982$.

In (41), R was a corporate bond rate (the yield on seasoned Moody's Aaa industrial bonds), M0 was the real per capita monetary base and y was real private GDP per capita. All the coefficients in (41) were significant and the mean lag for the impact of inflation on the interest rate was 8.14 quarters. These results are consistent with the liquidity preference theory.

Feldstein and Eckstein extended (41) to include privately held Federal government debt, and reported the estimates as:

$$R_t = -16.68 - 9.08 \ \ln \ \mathrm{M0}_t + 8.24 \ \ln \ y_t + 2.78D_t + 0.27\pi_t + \Sigma_j \alpha_j \pi_{t-j} \tag{42}$$

where $\Sigma_j \alpha_j = 3.93$, $j = 1, 2, \ldots, 23$, $R^2 = 0.985$, the mean lag for the inflation impact was 7.90 quarters and D was real per capita privately owned Federal government debt. In (42), the coefficient of government debt is positive since, as explained by the loanable funds theory, an increase in the supply of bonds, i.e. the demand for loanable funds, will raise the interest rate. In both (41) and (42) the mean lag for the impact of inflation on the interest rate is about 8 quarters and therefore quite long. Further, while the increase in the monetary base directly lowers the nominal interest rate, its indirect impact on inflation raises this rate. The reported coefficients imply that there does not exist short-term neutrality of money with respect to the real rate, at least not for periods up to 8 quarters.

While (41) represents the liquidity preference theory, (42) combines elements of both the liquidity preference and the loanable funds theories, the latter through its inclusion of

government debt. Since the coefficients of both the monetary base and government bonds are significant, the above estimates support the general version of the determination of the interest rate as given by a broader commodities–money–bonds model with Walras's law, rather than providing a rejection of either of the rival theories.

The general conclusion of the Feldstein and Eckstein (1970) study was that the rise in the interest rate between 1954 and 1965 was due more to decreasing liquidity than to inflation, but that inflation was more important from 1965 to 1969. The relatively slow growth of the public debt through the period held back the increase in the interest rate. Further, the direct impact of the changes in the monetary base and the government debt on the interest rate occurred within one quarter, so that there was not a significant lag in these effects, while the impact of inflation took place over 23 quarters, with a mean lag of about eight quarters. It is not clear whether such a long lag arises from a lag in the adjustment of the nominal interest rate to the expected rate of inflation, or from a lag in the expected rate in adjusting to the actual inflation rate. However, what is clear from this study is that Fisher's relationship between the interest rate and expected inflation must not be omitted in estimations of the nominal rate, even in studies based on the liquidity preference approach.

Sargent (1969), and Echols and Elliot (1976),[16] provided applications of the loanable funds theory. Sargent (1969) started with the identity:

$$R \equiv r^* + (r - r^*) + (R - r) \tag{43}$$[17]

where R is the nominal rate of interest (holding period yield on a bond). r is the real rate (the nominal rate less the expected rate of inflation over the holding period), and was called by Sargent the market real rate of interest. r^* is the rate of interest that equates investment and saving and corresponds to the "normal rate of interest" in Wicksell, as explained in Chapter 2. As in Wicksell's analysis, it was made a function of the excess of investment over saving. Its use by Sargent was meant to capture the loanable funds theory. $(r - r^*)$ is the deviation of the market real rate from the normal rate. As in Wicksell's analysis, this deviation depended upon the excess supply of money created through the bank's operations, which increases the supply of loans over that through saving. From the Fisher equation, $(R - r)$ equaled the expected rate of inflation in commodity prices. Sargent represented expectations by a distributed lag model.

The normal rate r^* was specified as a function of the excess demand for loans, which was defined as desired real investment less desired real saving. Investment was specified as a function of this rate and ΔX, while saving was a function of this rate and X, where X is real output. Among the reported estimates[18] are the following ones for the *ten-year bond yield*:

$$\begin{aligned} R_t = {} & 7.1338 + 0.0099\Delta X_t - 0.0456X_t - 2.0151(\Delta m^*{}_t / m^*{}_{t-1}) \\ & + 3.8764\ \Sigma_i\ 0.97^{i-1}(\Delta p_{t-i}/p_{t-i-1}) - 1.9849(0.97)^t \end{aligned} \tag{44}$$

Adjusted $R^2 = 0.9298$, $i = 1, 2, \ldots, t-1$.

16 The approach and results of this study will be examined in Chapter 20 in the context of the term structure of interest rates.

17 The symbols in this equation have been altered for consistency with our symbols in this chapter.

18 The Hildreth–Lu procedure was used to correct for serial correlation.

where R was the ten-year nominal bond yield, X was real GDP, m^* was real money supply and p was the commodity price index. All the coefficients were significant, except for ΔX_t, and the signs were consistent with the assumed hypotheses. An increase in the real money supply reduced the real rate of interest, with a 10 percent increase in the real money supply reducing the nominal interest rate by 20 basis points.

The estimated results for the *one-year bond yield* were:

$$R_t = 1.4396 + 0.0182\Delta X_t - 0.0405X_t - 6.0260(\Delta {m^*}_t/{m^*}_{t-1}) + 6.4716\ \Sigma_i\ (0.98)^{i-1}(\Delta p_{t-i}/p_{t-i-1}) + 4.4933(0.98)^t \tag{45}$$

Adjusted $R^2 = 0.9298$, $i = 1, 2, \ldots, t-1$.

In (45), the output level, changes in the money supply and the expectations variable were all significant and had the expected signs. The coefficient of the rate of change of output was insignificant in both (44) and (45). This variable was meant to capture the inclusion of the demand for loanable funds through investment.

Equations (44) and (45) show that both the money supply and the inflation rate have greater impact on the one-year rate than on the ten-year rate. Both indicate very long lags in the impact of inflation on the interest rate, possibly because of long lags in the process of expectations formation, though this result may have been due to the representation of expectations by a distributed lag function rather than by rational expectations.

Both equations include changes in the money supply, which is an element in the liquidity preference theory. Further, the inclusion of output could capture elements of money demand determination. Hence, it cannot be claimed that these equations exclude elements of liquidity preference. However, they also include elements of the traditional classical loanable funds theory. We tend to view them, as we did (41) and (42), as being consistent with the general macroeconomic model with Walras's law, and therefore with the interest rate moving in response to disequilibrium in both the money and bond markets.

Conclusions

This chapter has focused on the underlying interest rate in the economy, while leaving the study of the term and risk structure of interest rates to the next chapter. There are two main theories of this underlying rate. The loanable funds theory is associated with traditional classical economics and asserts that the interest rate is determined in the market for loanable funds – designated as bonds or credit in modern macroeconomic models. The liquidity preference theory is associated with Keynes and Keynesian economics, and asserts that the interest rate is determined by the equilibrium in the market for money.

In a completely specified macroeconomic model, the money and bond markets are only two of the markets in the economy. The other markets are those for commodities and labor. An interdependent structure of such a model implies that the interest rate is jointly determined with the other endogenous variables – including output and price level – of the model. For such a model, Walras's law implies that one of the markets can be omitted from explicit analysis. The loanable funds theory would omit the market for money while the liquidity preference theory would omit the market for bonds from explicit analysis, though both these choices would yield, *ceteris paribus*, the same equilibrium values of all the endogenous variables, including the interest rate. Hence, it does not matter for general equilibrium analysis which of these two theories is adopted in a given macroeconomic model.

Which theory is adopted does matter in a dynamic context. For analyzing the dynamic movements in the interest rate, our preference has been for the theory based on excess demand in the bond market.

However, empirical studies find support for the elements of excess demands for both money and bonds in explaining movement in the interest rate. An essential requirement for such empirical determination of the interest rate is the inclusion of the Fisher equation. While the empirical studies reported in this chapter used a distributed lag model for expectations and found long lags, newer studies tend to use rational expectations. No matter which procedure for modeling expectations is used, most studies report that increases in the money supply decrease the nominal interest rate and that money is not neutral in the short run as far as the nominal and real interest rates are concerned.

Summary of critical conclusions

- The traditional classical loanable funds theory stated that the real interest rate is determined by full-employment saving and investment in the economy.
- Keynes argued that there is no direct nexus between saving and investment. His liquidity preference theory asserted that the interest rate is determined by the demand and supply of money.
- The modern classical theory adapts the traditional classical loanable funds theory to the statement that the real interest rate is determined by the demand and supply of bonds.
- The modern classical approach implies that the long-run general equilibrium real rate of interest is invariant to anticipated changes in the money supply and the rate of inflation. Therefore, for the real rate of interest, there is neutrality of anticipated changes in money and inflation.
- The new Keynesian approach implies that monetary policy can change the real interest rate.
- Walras's law implies that it is immaterial in general equilibrium whether the money or the bond market is taken to be the proximate determinant of the rate of interest; the rate of interest will be identical.
- However, dynamic analysis shows that it does matter for the magnitude and in some cases also for the sign of the change in the interest rate whether this change is made a function of the excess demand for money or for bonds.
- In the modern financially developed economy, the more appropriate *proximate* determinant of changes in the interest rate is the excess demand for bonds. Changes in the excess demand for money cause changes indirectly in the interest rate by first changing the excess demand for bonds.

Review and discussion questions

1. Explain how the monetary and real factors enter into the determination of the interest rates in the short run and in the long run.
2. Compare and contrast the liquidity preference and loanable funds theories of the rate of interest. Discuss their implications for monetary policies intended to maintain full employment.
3. Keynes asserted that there is no such thing as a non-monetary theory of the rate of interest and that the rate of interest is uniquely determined by the demand and supply of money. Explain Keynes's reasons for this view. Compare this view with those of the traditional classical and modern classical schools.

4. Discuss the adjustment process likely to follow a change in (a) the money supply through open market operations, (b) a cut in the central bank's discount rate, leading to eventual changes in the interest rates in the economy.
5. Can the central bank change the interest rate in the economy through changes in its discount/bank rate? Present the analysis and theory relevant to your answer for the economy you live in.
6. Monetary theory implicitly assumes that the interest rates and the money stock are uniquely linked so that changes in one have a corresponding counterpart in the other, so that it does not matter which one the central bank chooses to change. What assumptions are needed for this assertion? Are they realistic enough for policy purposes?
7. This chapter has made the rather unusual assertion that in real-world economies there is no explicit market for money. Instead, the money market is a reflection of the other markets. Do you agree or disagree? Give reasons for your answer. What does your answer imply for the theory relevant to the determination of the interest rate if there is (a) general equilibrium in all markets, (b) disequilibrium in the economy?
8. The buffer stock analysis of the demand for money in Chapter 6 asserted that money acts as a buffer during periods in which economic agents need to adjust their stocks of other goods (commodities, bond and labor) to their optimal levels, but that such adjustments are more costly in the short term than those in money balances. What does this imply for the determination of the interest rate if there exists general equilibrium in all markets? What does it imply for the dynamic determination of the interest rate while there are buffer stock holdings of money following a shock that changes the desired demands for other goods?
9. Is there some relationship between the assertions (a) on buffer stock money holdings and (b) that in the real-world economies there is no explicit market in the economy for money but that it is a reflection of the other markets? Discuss.
10. Dynamic adjustments occur in disequilibrium but Chapter 18 raised doubts about the applicability of Walras's law if there was disequilibrium in the commodities and labor markets. In this context, should the dynamic analysis of interest rates be conducted with notional or effective excess demand functions? Discuss, keeping in mind that the objective is to explain the dynamic, disequilibrium determination of the rate of interest.
11. "The assumption of modern classical economics that there exists continuous labor market clearance at full employment means that we can confine the analysis of the real rate of interest to states of general equilibrium and ignore its properties for the disequilibrium states. Therefore, it does not matter whether the loanable funds theory or the liquidity preference theory was used: both imply the same rate of interest by virtue of Walras's law." Discuss the various aspects of this assertion.
12. Do the existence and operations of financial intermediaries have any implications for the rate of interest? If so, are these adequately reflected in the short-run macroeconomic models, and in what ways?
13. "The real rate of interest is a real variable. Under rational expectations, it is invariant to systematic changes in the money supply or the price level. However, unanticipated changes in these nominal variables can change the real rate of interest." Discuss this statement in the context of the neoclassical model and specify the implied Lucas-type equation for the determination of the real rate of interest. What are its implications for the pursuit of monetary policy?
14. Is there a "natural" rate of interest? What does it mean and what determines it? Is there a curve such as the Phillips curve for the real rate of interest? Discuss.

15. Why does the real interest rate fluctuate over the business cycle? Can monetary factors change it? Discuss.
16. Are the loanable funds and liquidity preference theories of the rate of interest consistent with (i) interest rate targeting, (ii) the Taylor rule? If not, how can they be made consistent?

References

Crowder, W.J., and Hoffman, D.L. "The long-run relationship between nominal interest rates and inflation: the Fisher equation revisited." *Journal of Money, Credit and Banking*, 28, 1996, pp. 102–18.

Echols, M.E., and Elliot, J.W. "Rational expectations in a disequilibrium model of the term structure." *American Economic Review*, 66, 1976, pp. 28–44.

Feldstein, M., and Eckstein, O. "The fundamental determinants of the interest rate." *Review of Economics and Statistics*, 52, 1970, pp. 363–75.

Hume, D. *Of Interest*. 1752. Reprinted in *The Philosophical Works of David Hume.* 4 vols. Boston: Little, Brown and Co., 1854. [Also available at: cepa.newschool.edu/het/profiles/hume.htm].

Keynes, J.M. *The General Theory of Employment, Interest and Money*. New York: Macmillan, 1936.

Mundell, R.A. "Inflation and real interest." *Journal of Political Economy*, 71, 1963, pp. 280–3.

Sargent, T.J. "Commodity price expectations and the interest rate." *Quarterly Journal of Economics*, 83, 1969, pp. 127–40.

Tobin, J. "Money and economic growth." *Econometrica*, 33, 1965, pp. 671–84.

20 The structure of interest rates

This chapter extends the determination of the single macroeconomic rate of interest to the multitude of interest rates in the economy.

Two of the major reasons for the variations among interest rates are the differences in the term to maturity and the differences in risk. To explain the former, it is important that the riskiness of bonds be held constant across assets of different maturities. This is made possible by confining the comparison to government bonds of different maturities and studying their yield curve. The main theory for explaining the term structure of interest rates is the expectations hypothesis.

Key concepts introduced in this chapter

- Yield curve
- Short rate of interest
- Long rate of interest
- Expectations hypothesis of interest rates
- Liquidity premium
- Segmented market hypothesis
- Preferred habitat hypothesis
- Random walk hypothesis

The short-run macroeconomic models of Chapters 13 to 15 have a single (bond) rate of interest, as analyzed in those chapters. However, there is more than one bond interest rate and more than one type of bond in the economy. By definition, the economist's concept of the rate of interest (or yield) on any given asset is the rate of return, including expected capital gains and losses, on that asset over a given period of time. Therefore, there is a rate of interest for each distinct type of asset in the economy. An example of this is provided by Chapter 16, which has two interest rates, one on bonds and the other on credit.

Assets differ in various aspects or characteristics. Some of the more significant differences consist in their marketability, their risk and their term to maturity. The rates of return on assets are likely to differ, depending upon their characteristics. The macroeconomic mode of focusing on only one rate of interest is quite acceptable if all interest rates

are related to each other in fixed proportions or fixed differences. Empirically, they do have a high positive correlation. The relationship between prices and rates of return on assets of differing maturities is brought out by the theories on the term structure of interest rates. These theories and the empirical work based on them are the focus of this chapter.

Section 20.1 defines the spot, forward and long rates of interest. Section 20.2 sets out the theories explaining the term structure of interest rates. Of these, the most significant one for developed financial markets is the expectations hypothesis. Section 20.3 briefly touches on the relationship between asset prices and yields, and on tests based on the term structure of asset prices. Sections 20.4 and 20.5 report on some of the empirical work on the term structure of interest rates. Section 20.6 presents the random walk hypothesis which is related to the expectations hypothesis and uses the rational expectations hypothesis for the formation of expectations. Section 20.7 uses the term structure to derive estimates of the expected rate of inflation.

The basic model of the relationship between the prices and yields of assets with different risks is the capital asset pricing model proposed by Sharpe (1964). This analysis is based on the expected utility hypothesis developed in Chapter 5. However, such analysis is usually not included in textbooks on monetary economics and, for reasons of brevity, we have chosen not to include it in this book.

Note that the theories on the risk structure and the term structure of interest rates only explain the interest rate differentials due to differences in risk or the term to maturity, and do not explain the basic interest rate in the economy, which was the subject of Chapter 19.

Notation

Unfortunately, the notation in this chapter has to be quite cumbersome, so that some explanation on its general pattern would be useful. To start, first note that all interest rates in this chapter are nominal. The *short* (nominal, one-period) *interest rates* are designated by r.[1] The current period is designated as t. Suppose that a contract is entered into in period $t+j$ for a one-period loan for period $t+i$ at an interest rate r, $j \leq i$. This will be written as ${}_{t+j}r_{t+i}$, where the left subscript indicates the period in which the contract is made and the right subscript indicates the period for which the loan is made. If future interest rates are the expected ones, we would write the corresponding rate as ${}_{t+j}r^{e}_{t+i}$ and its rational expectation as $E_{t+j}\,{}_{t+j}r_{t+i}$, where the expectation E_{t+j} is based on information available in period $t+j$.

The *long rates* are designated by the capital symbol R. The contract for these is always assumed to be entered into in the current period t. ${}_{t}R_{t+i}$ will designate the long rate on a contract for a loan of i periods. Since this interest rate is known in the current period t, it is an actual rather than an expected rate.

20.1 Some of the concepts of the rate of interest

The short-term markets for bonds have spot, forward and long rates of interest. The meanings of these terms are as follows.

1 Note that the lower-case symbol r designates a nominal short rate. In earlier chapters, the nominal rates were designated by the capital symbol R. However, in this chapter, R will be reserved for the nominal long interest rate.

The (current) spot rate of interest

The (current) spot rate of interest ${}_tr_t$, or written simply as r_t, is the annualized rate of return on a loan for the current period t, with the loan being made at the beginning of period t.

The future spot rate of interest

The future spot rate of interest is the return on a one-period loan in a future period $(t+i)$, $i>0$, with the loan made at the beginning of that period. It will be designated ${}_{t+i}r_{t+i}$ or r_{t+i}, so that the left-hand subscript will be implicit. Since r_{t+i} is a future spot rate, its expected value will be designated $r_{t+i}{}^{\mathrm{e}}$. Its rational expectation in period t will, then, be written as $E_t r_{t+i}$, or as $E_{t\,t+i}r_{t+i}$.

The future short rate of interest

The future short rate of interest is the return on a one-period loan in a future period $(t+i), i>0$, with the contract for the loan entered into at the beginning of period $t+j$, $j \le i$, which could be the current period. It will be designated ${}_{t+j}r_{t+i}$.

The forward short rate of interest

The forward short rate of interest ${}_tr^{\mathrm{f}}{}_{t+i}$ is the annualized rate of interest on a one-period loan for the $(t+i)$th period only, with the contract for the loan being made in the current period t. Note that the superscript f has been inserted to stand for "forward." The forward rate differs from the future short rate ${}_{t+i}r_{t+i}$ (or r_{t+i}), where the one-period loan for the period $(t+i)$ is contracted at the beginning of period $t+i$. In incomplete financial markets, ${}_tr^{\mathrm{f}}{}_{t+i}$ may not exist but ${}_{t+i}r_{t+i}$ would do so as long as there are spot markets. However, ${}_tr^{\mathrm{f}}{}_{t+i}$, if it exists, will be known in the current period t, whereas ${}_{t+i}r_{t+i}$ is not likely to be known in t, though expectations on its value can be formed in t.

The long rate of interest

The long rate of interest ${}_tR_{t+i}, i=0,1,\ldots,n$, is the rate of return per period on a loan for $(i+1)$ periods, the loan being made in period t, with repayment of the principal and accumulated interest after $(i+1)$ periods.

The current spot rate of interest ${}_tr_t$ and the one-period long rate of interest ${}_tR_t$ are identical. For simplicity of notation, ${}_tr_{t+i}$ will sometimes be written as r_i and ${}_tR_{t+i}$ will be written as R_i, with the subscript t being implicit or with the current period being treated as 0.

20.2 Term structure of interest rates

20.2.1 Yield curve

The variation in yields on assets of different maturities (redemption dates) is known as the term structure of interest rates, with the assets being assumed to be identical in all respects except for their maturity. This requirement is generally fulfilled only by the bonds issued by the government, so that the yields on government bonds are examined to show the variation in yield with increasing maturity. This variation is shown graphically by plotting the nominal

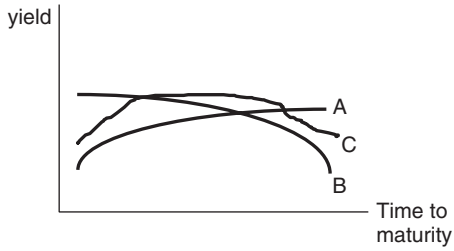


Figure 20.1

yield r on government bonds on the vertical axis and the time up to maturity on the horizontal axis, as in Figure 20.1. The curve thus plotted is known as the *yield curve*.

The yield curve normally slopes upward from left to right, with the yield rising with term to maturity, as shown by the curve A in Figure 20.1. It can, however, possess any shape. In times of monetary stringency, short-term interest rates can rise and move above the long-term rates, as shown by curve B. This can also happen when inflation is rampant in the economy but is expected to be a short-term problem so that the inflationary premium in nominal yields is greater for the shorter term than for the longer term bonds. In some cases, the curve may have a hump, as shown by curve C. In this case, some intermediate securities have the highest yield, usually because of the expectation that the highest rates of inflation will occur in the intermediate periods.

The two main determinants of the shape of the yield curve in practice are the time structure of the expected inflation rates and the current stage of the business cycle. On the former, as explained in several earlier chapters, Fisher's relationship between the nominal yields and the expected inflation rate is:

$$(1 + r_t) = (1 + r^{\mathrm{r}}{}_t) + (1 + \pi^{\mathrm{e}}{}_t)$$

where r is now the nominal short yield, r^{r} is the real short yield and π^{e} is the expected inflation rate. The higher the expected rate of inflation, the more will the time structure of expected inflation determine the shape of the yield curve.

The yield curve changes its shape over the business cycle. Long-term yields are usually higher than short-term yields mainly because long-term debt is less liquid and is subject to greater price uncertainty than short-term debt. However, the short-term yields are more volatile, rising faster and extending further than long yields during business expansions and falling more rapidly during recessions. Large swings in short-term rates, and to a lesser extent in intermediate rates, together with relatively narrow movements in long-term rates, cause a change in the shape of the yield curve over the course of a business cycle.

A sharp increase in short-term rates frequently occurs near the peak of a business expansion because of a combination of factors, most often including a strong demand for short-term credit, restrictive effects of monetary policies on the supply of credit, and changing investor expectations. Depending upon the intensity of these forces, the yield curve will be relatively flat, have a slight downward slope, or show a steep negative slope. As short rates fall absolutely and relative to long yields during the ensuing economic slowdown, the yield curve tends to regain its positive slope, acquiring its steepest slope near the cyclical trough. As the economy recovers and economic activity picks up, short rates again rise faster than long yields, and the

yield curve tends to acquire a more moderate slope. Since the yield curve plots the nominal rather than the real rate of interest, and the nominal rate includes the expected rate of inflation, the dominant element of the shape and shifts in the yield curve is often the term structure of the expected rate of inflation.

There are basically three main theories on the term structure of interest rates. These are:

1 The expectations hypothesis, first formulated by Irving Fisher. This theory is the relevant one for financially developed markets, and is supported by most empirical studies.
2 The segmented markets theory, with Culbertson as its major proponent.
3 The preferred habitat hypothesis.

20.2.2 Expectations hypothesis

Irving Fisher in *The Theory of Interest* (1930, pp. 399–451) considered the rate of return or yields on securities that differ only in terms of their maturity. His approach assumes that:

(i) All borrowers and lenders have perfect foresight and know future interest rates and asset prices with certainty, so that there is no risk. An alternative assumption to this is that, while there is uncertainty of yields, the borrowers and lenders are risk neutral and form rational expectations about the future short rates.
(ii) There are no transactions costs in switching from money into securities and vice versa.
(iii) The financial markets are *efficient*.

A market is said to be *efficient* if it clears (i.e. demand equals supply) instantly and prices reflect all available information. In such a market, any opportunities for superior profits are instantly eliminated. By comparison, a *perfect market* assumes perfect competition among traders *and* efficient markets. Fisher's assumptions specify an efficient market, which need not have perfect competition, so that it need not be a perfect one.

Investors are assumed to maximize their expected utility, subject to the relevant constraints. However, under assumption (i), this is synonymous with the maximization of the expected return to the portfolio. Under assumptions (i) and (ii), a lender wishing to make a loan for n periods will be indifferent between an n-period loan or a succession of n one-period loans only if the overall return were the same in both cases. Under assumption (iii), with all investors acting on this basis, the market yields will be such as to ensure this indifference.

Expectations hypothesis, complete markets and forward rates

Assume that the financial sector has complete markets, so that there exist markets for long loans of all possible maturities, as well as for spot and *forward* one-period loans. With the current period as t, the yield (per period) on an $(i+1)$-*period* loan was designated as ${}_tR_{t+i}$, while that on a *one-period* loan for the $(i+1)$th period was ${}_tr^{\mathrm{f}}_{t+i}$, $i = 0, 1, \ldots, n$, where $n+1$ is the longest maturity available in the market. Hence, ${}_tr_t$ is the (spot) yield on a loan for the first period; ${}_tr^{\mathrm{f}}{}_{t+1}$ is the forward yield on a loan for the second period; and so on. An $(i+1)$-*period* loan of \$1 will pay the lender $(1+{}_tR_{t+i})^{i+1}$ at the end of the $(i+1)$th period. The series of $(i+1)$ loans starting with a principal of \$1 for one-period at a time will pay him $[(1+{}_tr_t)\,(1+{}_tr^{\mathrm{f}}{}_{t+1})\ldots(1+{}_tr^{\mathrm{f}}{}_{t+i})]$ at the end of the $(i+1)$th period. Under the above three assumptions, the lender will be indifferent between the two types of loans if the total amount

repaid to him after $n+1$ periods is identical. With all investors exhibiting this behavior, efficient markets under certainty ensure that:

$$(1+{}_tR_{t+i})^{i+1} = (1+{}_tr_t)(1+{}_tr^{\mathrm{f}}{}_{t+1})(1+{}_tr^{\mathrm{f}}{}_{t+2}) \ldots (1+{}_tr^{\mathrm{f}}{}_{t+i}) \tag{1}$$

This formula will hold for every $i, i = 0, \ldots, n$, where $n+1$ is the longest maturity in the market, so that:

$$(1+{}_tR_t) = (1+{}_tr_t)$$

$$(1+{}_tR_{t+1})^2 = (1+{}_tr_t)(1+{}_tr^{\mathrm{f}}{}_{t+1})$$

$$(1+{}_tR_{t+2})^3 = (1+{}_tr_t)(1+{}_tr^{\mathrm{f}}{}_{t+1})(1+{}_tr^{\mathrm{f}}{}_{t+2})$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$(1+{}_tR_{t+n})^{n+1} = (1+{}_tr_t)(1+{}_tr^{\mathrm{f}}{}_{t+1})(1+{}_tr^{\mathrm{f}}{}_{t+2}) \ldots (1+{}_tr^{\mathrm{f}}{}_{t+n}) \tag{2}$$

Under our assumption of complete markets, the forward rates are known, rather than merely expected, in period t. However, even well developed financial markets do not have forward markets for all future periods, so that (2) cannot be applied for all maturities.

Expectations hypothesis and expected future spot rates

Since there would always be spot markets over time, designate the spot rate expected in period t for the period $t+i$ as ${}_tr^{\mathrm{e}}{}_{t+i}$, where the subscript t on the left side in the presence of the superscript e indicates that the expectations are formed in period t for the spot rate for period $t+i$.[2] The investor would then have a choice of investing long for $t+i$ periods, with a known long rate ${}_tR_{t+i}$, and investing over time in a sequence of spot markets at the spot rates in those markets. In practice, since these future spot rates can differ from the actual ones, there is a risk in following the latter strategy. The investor will be indifferent between the two strategies if he is *risk indifferent* and if their expected return is identical. Hence, in terms of the expected future rates, the expectations hypothesis becomes:

$$(1+{}_tR_{t+i})^{i+1} = (1+{}_tr_t)(1+{}_tr^{\mathrm{e}}{}_{t+1})(1+{}_tr^{\mathrm{e}}{}_{t+2}) \ldots (1+{}_tr^{\mathrm{e}}{}_{t+i}) \tag{3}$$

Note that (3) differs from (1) since (3) involves expected future spot rates while (1) involves the corresponding forward rates, which are known in period t. For many investors, though ones with relatively small portfolios, the assumptions of the expectations hypothesis can be somewhat unrealistic. There is often both a transfer cost in and out of securities and a lack of perfect foresight (or risk indifference) about the future. The former implies that n one-period loans will involve much greater expense and inconvenience than a single n-period loan. The latter implies that loans of different maturities involve different risks and, for risk averters, a higher risk has to be compensated for by a higher yield. For very many large transactors, usually financial institutions, the transactions costs tend to be negligible, so that (3) should hold approximately, if not accurately.

2 Note that ${}_tr^{\mathrm{f}}{}_{t+i}$ is a forward rate contracted in t for the one-period loan in $t+i$, whereas ${}_tr^{\mathrm{e}}{}_{t+i}$ is the spot rate for $t+i$ with expectations formed in t.

Under the rational expectations hypothesis, r^e is replaced by $E_t r$, so that (3) becomes:

$$(1 +_t R_{t+i})^{i+1} = (1 +_t r_t)(1 + E_{t\,t} r_{t+1})(1 + E_{t\,t} r_{t+2}) \ldots (1 + E_{t\,t} r_{t+i}) \tag{3'}$$

If a difference emerges in the markets between the left and the right sides of (1) and (3), profits can be made through arbitrage, which would take place to establish their equality. The rest of this chapter proceeds in terms of (3) or (3′) rather than (1). While financial markets, even in developed economies, rarely have a large number of forward markets, they usually do have markets for government securities of many different maturities. The long rates of interest are quoted on these securities, so that their values are known each period. These values can be used to calculate the expected short rates of interest by using the following iterative reformulation of (3):

$$\begin{aligned} E_{t\,t} r_{t+1} &= (1 +_t R_{t+1})^2 / (1 +_t r_t) - 1 \\ E_{t\,t} r_{t+2} &= (1 +_t R_{t+2})^3 / [(1 +_t r_t)(1 + E_{t\,t} r_{t+2})] - 1 \end{aligned} \tag{4}$$

and so on.

If the market forms its expectations in terms of the expected future short rates, the long rates will be determined from these short rates by the preceding equations. Some economists assume that the investors' expectations are formed in terms of a series of expected short rates for the future periods, while others assume that investors are concerned with the prices of the assets currently in the market and that these prices can be used to calculate the long rates. Therefore, equation (3) can be used from right to left or from left to right.

Long rates as geometric averages of short rates

According to (3), the long rates are geometric averages of the short rates of interest. This implies that:

1 If the short interest rates are expected to be identical, the long rate will equal the short rate.
2 If the short interest rates are expected to rise, the long rates will lie above the current short rates.
3 If the short interest rates are expected to decline, the long rates will be less than the current short rate.
4 The long rate, being an average of the short rates, will fluctuate less than the short rate.

In principle, any pattern of expected future short rates is possible, with the result that some long rates may be less and some greater than the current spot rate, so that the yield curve may have any shape whatever.

The assumptions of the expectations hypothesis may not always hold for all agents in the market, which encompasses both households and firms. However, developed financial economies tend to be competitive and efficient. Therefore, the expectations hypothesis will hold if the credit markets have sufficient numbers of participants who behave according to the assumptions of perfect foresight (or of rational expectations and risk indifference) and zero variable transfer costs between securities and money. These assumptions tend to be valid at least for large financial institutions operating in the developed economies. Hence, the expectations hypothesis should be more or less valid for developed financial markets.

Once the market has established a structure of short and long rates according to (1) or (3), the demand and supply functions for long and short bonds on an individual basis will become *indeterminate*: an investor would be indifferent between a long bond maturing at the end of the *i*th period and various sequences of short and long bonds with a corresponding combined maturity.

20.2.3 Liquidity preference version of the expectations hypothesis

Both the *n*-period loan and a series of *n* one-period loans involve risks, though of different kinds. The *n* one-period loans involve the possibility that the future spot rates will turn out to be lower than the expected forward rates or the *n*-period long rate. This is an *income* loss. But the *n*-period loan – that is, purchase of a bond maturing after n periods – involves the possibility that the lender may need his funds somewhat sooner and have to sell the bond before it matures. Such a sale may involve a *capital* loss, especially in the absence of a secondary market for loans. There is also the possibility that more profitable opportunities may turn up and have to be foregone if the funds are already loaned up for a long period.

It is likely that the possibility of a capital loss influences lenders' decisions more than that of the interest loss since the capital loss can usually take on much greater magnitude than the interest loss. Further, if the funds represent precautionary saving, the individual would prefer a more liquid (shorter maturity) to a less liquid (longer maturity) asset. Hicks (1946, pp. 151–82) suggested that lenders wish to avoid the risk of a capital loss by investing for shorter rather than longer periods. Therefore, under uncertainty of future yields, they have to be compensated by a higher yield on longer term loans. Conversely, borrowers – generally firms borrowing for long-term investments – prefer borrowing for a longer term than for a shorter term, which makes them willing to pay a premium on longer term loans. Such risk-avoidance behavior on the part of both lenders and borrowers implies that the longer term loans will carry a premium over shorter term loans. Hence, the yield on bonds will increase with the term to maturity, so that equation (3) will be modified to:

$$(1 + {}_tR_{t+n})^{n+1} > (1 + {}_tr_t)(1 + {}_tr^{e}{}_{t+1})\ldots(1 + {}_tr^{e}{}_{t+n}) \qquad n \geq 1 \tag{5}$$

Equation (5) is known as the *liquidity preference hypothesis of the yield curve*. For a more specific hypothesis on liquidity preference, designating the *liquidity premium* as ${}_t\gamma_{t+n}$, we have:

$$(1 + {}_tR_{t+n})^{n+1} = (1 + {}_tr_t)(1 + {}_tr^{e}{}_{t+1})\ldots(1 + {}_tr^{e}{}_{t+n})\,{}_t\gamma_{t+n}(n;\rho) \qquad n \geq 1 \tag{6}$$

where $\partial\gamma_n/\partial n \geq 0$ by virtue of the liquidity premium, and:

γ = liquidity premium
ρ = degree of risk aversion
n = periods to maturity.

We can distinguish between two versions of (6) on the basis of two alternative assumptions on the liquidity premium. These are that:

(i) The liquidity premium is constant at γ per period, so that ${}_t\gamma_{t+i} = i\gamma$. While there is no particular intuitive justification for making this assumption, it is analytically convenient and, as seen later in this chapter, is made in many empirical studies. It reduces (6) to:

$$(1 + {}_tR_{t+n})^{n+1} = (1 + {}_tr_t)(1 + {}_tr^{e}{}_{t+1})\ldots(1 + {}_tr^{e}{}_{t+n})n\gamma \qquad n \geq 1 \tag{7}$$

Equation (7) with a constant per period risk premium is sometimes called the *strong form* of the expectations hypothesis with a liquidity premium.

(ii) The per period liquidity premium varies with the term to maturity and, moreover, may not be constant over time, e.g. over the business cycle, so that (6) does not simplify to (7). This is sometimes called the *weak* form of the expectations hypothesis with a liquidity premium. Estimation of this form requires specification of the determinants of the liquidity premium.

Compared with these *weak* and *strong* forms of the expectations hypotheses, the original form (3) of this hypothesis without a liquidity premium is known as the *pure* form of the expectations hypothesis.

20.2.4 Segmented markets hypothesis

If the uncertainty in the loan market is extremely severe or if lenders and borrowers have extremely high risk aversion, each lender will attempt to lend for the exact period for which he has spare funds and each borrower will borrow for the exact period for which he needs funds. In this extreme case, the overall credit market will be split into a series of segments or separate markets based on the maturity of loans, without any substitution by either borrowers or lenders among the different markets. Therefore, the yields in any one market for a given maturity cannot influence the yields in another market for another maturity. Hence, there would not be any particular relationship such as (3) or (6) between the long and the short rates, and the yield curve could have any shape whatever. This is the basic element of the segmented markets theory: the market is segmented into a set of independent markets. Culbertson (1957) stressed this possibility as a major, though not the only, determinant of the term structure of interest rates.

Culbertson also argued that the lender rarely knows in advance exactly when he will need his funds again and will prefer to make loans for shorter terms rather than longer ones, the former being the more liquid of the two. If the supply of short-term debt instruments is not sufficient to meet this demand for liquidity at a rate of interest equal to the long-term rate, the short-term rate will be less than the long-term rate. Further, the supply of short-term instruments is generally limited since lenders will not finance long-term investment with short-term borrowing. Therefore, the short-term yield will be less than the long-term yield, *ceteris paribus*.

The segmented markets hypothesis is more likely to be applicable in the absence of developed financial markets, including secondary markets for securities, and sophisticated investors. It may therefore be somewhat more valid for financially underdeveloped markets than for developed ones.

20.2.5 Preferred habitat hypothesis

The preferred habitat hypothesis was proposed by Modigliani and Sutch (1966, 1967), and represents a compromise between the expectations hypothesis of perfect substitutability and the segmented markets hypothesis of zero substitution between loans of different maturities. Modigliani and Sutch argued that lenders would prefer to lend for periods for which they can spare the funds and borrowers would prefer to borrow the funds for periods for which they need the funds. However, each would be willing to substitute other maturities, depending upon their willingness to take risks and the opportunities

provided by the market to transfer easily among different maturities. Bonds maturing close together would usually be fairly good substitutes and have similar risk premiums. This would be especially so for bonds at the longer end of the maturity spectrum. Therefore, in well developed financial markets, a high degree of substitutability would exist among different maturities, but without these necessarily becoming perfect substitutes. Hence, while the yields on different maturities would be interrelated to a considerable extent, there would also continue to exist some variation in yields among the different maturities.

20.2.6 Implications of the term structure hypotheses for monetary policy

The expectations theory and the segmented markets theory have significantly different implications for the management of the public debt and for the operation of monetary policy. The expectations theory implies that the market substitution between bonds of different maturities is so great that a shift from short-term to long-term borrowing by the government will not affect the shape of the yield curve. The segmented markets theory implies that a substantial purchase (sale) of short-term bonds will lower (raise) the short-term interest rates while a sale (purchase) of long-term bonds will raise (lower) the long-term rates, so that such policies can alter the yield curve. The implications of the preferred habitat hypothesis lie between those of the expectations hypothesis and the segmented markets hypothesis, and are closer to one or the other depending upon the stage of development of the financial markets and the characteristics of the economic agents operating in them.

The empirical evidence for economies with well-developed financial markets has so far generally favored the expectations theory or a version of the preferred habitat hypothesis close to the expectations hypothesis over the segmented markets hypothesis. Intuitively, the credit markets for such economies are not seriously segmented since borrowers and lenders do generally substitute extensively between assets of different maturities.[3] A number of studies for the USA and Canada have substantiated the expectations theory at the general level, though there also exist many empirical studies that reject its more specific formulations. We discuss some of these later in this chapter.

20.3 Financial asset prices

Financial assets are not generally held for their direct contribution to the individual's consumption. They are held for their yield, which is often uncertain, and the individual balances the expected yield against the risks involved. This is the basic approach of the theories of portfolio selection. These theories focus on the yields on assets rather than on the prices of assets.

The price of any asset is uniquely related to its yield and can be calculated from the following relationship. In any period t, for an asset j,

$$r_{jt} = ({}_{t}p^{e}{}_{jt+1} - p_{jt}) + x_{jt} \tag{8}$$

3 An early study by Meiselman in 1962 supported the expectations theory. He also found that there was not sufficient justification for the assumption of a liquidity premium.

where:

r_{jt} = expected yield on the jth asset during period t
p_{jt} = jth asset's price in period t
${}_tp^e{}_{j\,t+1}$ = jth asset's (expected) price in period $t+1$, with expectations held in t
x_{jt} = jth asset's coupon rate in period t.

That is,

$$ {}_tp^e{}_{jt+1} = p_{jt} + r_{jt} - x_{jt} \tag{9} $$

Hence, a theory of the rate of interest is also a theory of the prices of financial assets. Alternatively, the yields on assets may be explained by a theory of asset prices. Such a theory at a microeconomic level would consider the market for each asset, and use the demand and supply functions for each asset to find the equilibrium price of the asset. At the macroeconomic level, the theory could focus on the average price of financial assets, with macroeconomic demand and supply functions. These demand and supply functions would have the prices of the assets as the relevant variables. This suggests two structural estimation procedures. One of these would specify the demand and supply functions for financial assets in terms of their prices, while the second one would do so in terms of interest rates. The former procedure would derive the equilibrium prices of assets with different maturities, which can then be used to calculate the short and long interest rates. The latter procedure would derive the equilibrium short and long interest rates, which can be used to calculate the prices of assets of different maturities. These procedures are not explicitly specified in this book but can be found in its first edition (2000). Empirical studies based on these structural approaches include, among others, Benjamin Friedman (1977), Feldstein and Eckstein (1970), Sargent (1969) and Echols and Elliot (1976).

An illustration of the arguments and findings of such studies is provided by Echols and Elliot (1976). These authors extended Sargent's (1969) analysis and tested for the determinants of forward rates using real GNP, government deficit, net export balance, real money supply, stock of outstanding government bonds, bank funds and insurance company funds invested in government bonds – as well as inflationary expectations – among their explanatory variables. Their estimations of forward rates for US data found the coefficients of the explanatory variables to be significant, with signs consistent with the loanable funds approach. Among their results was the significance of the liquidity premium, as well as that of the institutional (bank and insurance company) demands for bonds of different maturities and the supply of short versus long maturity supplies of government bonds, thereby supporting the preferred habitat hypothesis.[4] For example, an increase in the proportion of investment funds held by banks relative to insurance companies lowered forward rates. However, these institutional holdings and supply factors did not prove to be significant in explaining the yield spread between twelve-year government bonds and Treasury bills, so that it is not clear that the government could shift the yield curve by debt management policy – e.g. by increasing the issue of short-term government bonds relative to long-term bonds.

Compared with these structural approaches, the ones usually employed to test the theories of the term structure of interest rates are reduced-form approaches. These are based on the expectations hypothesis and are considered in detail below.

4 These results of Echols and Elliot (1976) should be compared with those in Pesando (1978), discussed later in this chapter.

20.4 Empirical estimation and tests

20.4.1 Reduced-form approaches to the estimation of the term structure of yields

As shown earlier, the expectations hypothesis modified with the addition of a liquidity preference term implied (6), which was that:

$$(1 + {}_tR_{t+n})^{n+1} = [(1 + {}_tr_t)(1 + {}_tr^e{}_{t+1})\ldots(1 + {}_tr^e{}_{t+n})]_t\gamma_{t+n}(n;\rho) \qquad n \geq 1 \qquad (10)$$

where the current period is t, the expectations are those held in period t and γ represents the liquidity premium, which depends on the term to maturity n and the risk premium ρ. Now, using the symbols R and r for the *logarithmic* values of the *gross* (rather than the net) returns and adding a random error η_t, the estimation form of (10) becomes:

$${}_tR_{t+n} = \{1/(n+1)\}[{}_tr_t + {}_tr^e_{t+1} + \ldots + {}_tr^e_{t+n} + {}_t\,\gamma_{t+n}(n;\rho)] + \eta_t \qquad n \geq 1 \qquad (11)^5$$

where *all the return variables are gross rates of return and all symbols now indicate log values*. We will follow this convention in the rest of this chapter.

In order to test (11), we need a hypothesis for generating the expected values of the forward short rates. Given the efficient markets assumption, the natural complement of the expectations hypothesis is the rational expectations hypothesis (REH) presented in Chapters 8 and 14. This hypothesis specifies that:

$${}_tr^e{}_{t+i} = E_t r_{t+i} \qquad (12)$$

where $E_t\ r_{t+i}$ is the rational expectations value of r_{t+i} based on all the information available in t about period $t+i$. Among the information needed is that of the "relevant theory" that actually determines r_{t+i} and information on the values of its determinants. This throws us back to the demand and supply functions for assets for the $(t+i)$th period. Alternatively, if we can assume that this theory and the values of the explanatory variables are all known, we can estimate the stochastic equation (11).

As an illustration of this point, assume that the *relevant theory* is the simple autoregressive relationship with a stochastic term:

$$r_{t+i+1} = a_1 r_{t+i} + a_2 r_{t+i-1} + \mu_{t+i+1} \qquad i = 0, 1, 2, ..., n \qquad (13)$$

where μ_t is a random error with zero mean and constant variance. Under the REH,

$$E_t r_{t+i+1} = a_1 E_t r_{t+i} + a_2 E_t r_{t+i-1} \qquad (14)$$

where $E_t\ r_{t+i}$ is the rational expectation of the expected spot rate ${}_tr^e{}_{t+i}$, with the expectations formed in t. By iteration, $E_t r_{t+i}$ can be expressed as functions of r_t and r_{t-1}, whose values are already known in t. These, along with the expectations hypothesis of the term structure, can be used to generate $E_t r_{t+i}$ for all i. The following provides an example of these arguments for $i = 2$.

5 Some researchers have called (11) the "*fundamental equation*" of the term structure and bond pricing.

For our example, (11) for i = 2 (that is, three period) becomes:

$$ {}_tR_{t+2} = (1/3)[r_t + {}_tr^e{}_{t+1} + {}_tr^e{}_{t+2} + {}_t\gamma_{t+2}] + \eta_t \tag{15} $$

Combining (15) with the REH, we get:

$$ E_tR_{t+2} = (1/3)[r_t + E_tr_{t+1} + E_tr_{t+2} + {}_t\gamma_{t+2}] \tag{16} $$

where $E_t R_{t+2}$ is the mathematical expectation in t of the long rate ${}_tR_{t+2}$ from t to $t+2$. From (16) and (14), we have:

$$ \begin{aligned} E_tR_{t+2} &= (1/3)[r_t + (a_1r_t + a_2r_{t-1}) + ({a_1}^2 + a_2)r_t + a_1a_2r_{t-1}] + {}_t\gamma_{t+2} \\ &= \alpha_1 r_t + \alpha_2 r_{t-1} + (1/3){}_t\gamma_{t+2} \end{aligned} \tag{17} $$

where $\alpha_1 = (1/3)(1 + a_1 + a_2 + {a_1}^2)$ and $\alpha_2 = (1/3)(a_2 + a_1a_2)$. Since the REH implies that:

$$ {}_tR_{t+2} = E_tR_{t+2} + \eta_t \tag{18} $$

we have from (17) and (18) that:

$$ {}_tR_{t+2} = \alpha_1 r_t + \alpha_2 r_{t-1} + (1/3){}_t\gamma_{t+2} + \eta_t \tag{19} $$

(19) is the estimating equation given the expectations hypothesis, the REH and the assumed specification of the relevant theory as (13). Its general form is:

$$ {}_tR_{t+i} = \alpha_1' r_t + \alpha_2' r_{t-1} + (1/(i+1)){}_t\gamma_{t+i} + \eta_t \tag{20} $$

for appropriate definitions of α_1' and α_2' in terms of a_i. Their estimated values would reflect the influence of the three underlying hypotheses. However, note that (20) requires data on the risk/liquidity premium ${}_t\gamma_{t+i}$, which is not observable, so a hypothesis on it will have to be specified before (20) can be estimated. The usual assumptions on ${}_t\gamma_{t+i}$ are considered next.

Two common hypotheses on the risk premium

The simplest possible hypotheses on the risk/liquidity premium ${}_t\gamma_{t+i}$ are:

(i) ${}_t\gamma_{t+i}$ is constant per period such that ${}_t\gamma_{t+i} = i\gamma$, where the liquidity premium for $(i+1)$ periods involves this premium for only i periods (after the current one). In this case, this term will become the constant in (20).

(ii) ${}_t\gamma_{t+i}$ is random such that ${}_t\gamma_{t+i} = \xi_{t+i}$. In this case, the liquidity term will become part of the random term in (20).

(i) is the more common assumption in the estimation of (20) and is used in the next section.

20.5 Tests of the expectations hypothesis with a constant premium and rational expectations

There is no particular basis for assuming that the liquidity premium is constant. However, making such an assumption facilitates the construction of empirical tests of the

expectations hypothesis. The following two tests are based on this assumption. These tests use the implications of the expectations hypothesis with a constant liquidity premium, so that the expected (holding period) yields on the relevant bonds of different maturities will differ only by a constant representing the liquidity premium.

Define the difference between the *actual* long yield and the *average* one specified by the right side of the expectations equation (1) as the excess yield on the long bond. From (11), the actual difference in the yields from holding a long bond as opposed to a sequence of short bonds is related to the liquidity premium and expectational errors. Assuming this premium to be constant and assuming rational expectations, the remaining variations in the excess yields can then only be due solely to random fluctuations. This, in turn, implies that the difference between the excess yield and the premium will be due solely to random errors in expectations and cannot be forecast with any information known at the time the expectations are formed.

20.5.1 Slope sensitivity test

For this test, start with the following non-stochastic form for the *two*-period long rate:

$$2\,{}_tR_{t+1} = {}_tr_t + {}_{t+1}r^{e}{}_{t+1} + {}_t\gamma_{t+1} \qquad t = 0, 1, \ldots \tag{21}$$

where, as a reminder, note that all the variables are in logs and the interest rate variables are gross rates. Assuming the constancy of the liquidity premium per period, and noting that a two-period loan involves a liquidity premium only for the second period, let:

$${}_t\gamma_{t+1} = \gamma \tag{22}$$

(21) can be restated as:

$${}_{t+1}r^{e}{}_{t+1} - {}_tr_t = 2[{}_tR_{t+1} - {}_tr_t] - \gamma \tag{23}$$

Assuming the rational expectations hypothesis,

$${}_{t+1}r^{e}{}_{t+1} = E_{t\,t+1}r_{t+1} \tag{24}$$

and

$${}_{t+1}r_{t+1} = E_{t\,t+1}r_{t+1} + \mu_{t+1} \tag{25}$$

where μ_t is a random error with $E_t(\mu_{t+1}|I_t) = 0$ and I_t is the information available in t. Hence, from (23) to (25),

$${}_{t+1}r_{t+1} - {}_tr_t = \alpha + \beta({}_tR_{t+1} - {}_tr_t) - \mu_{t+1} \tag{26}$$

where $\alpha = -\gamma$ and $\beta = 2$. Since each of the variables (except for the random term) in (26) is observable, it can be estimated by the appropriate regression technique. This equation specifies that the change over time in the one-period spot rates from the current to the next period will depend upon the difference between the current two-period long rate and the current one-period spot rate, except for a constant term and a random term. New information

appearing in the next period will do so randomly. If the regressions of the above equation yield estimates consistent with these restrictions on α, β and μ, the theory will not be rejected by the data.

Equation (26) provides a joint test of three hypotheses: the expectations hypothesis, the REH and a constant liquidity premium per period. Since the test is that the estimated value of β does not significantly differ from 2, it is called the *slope sensitivity test*.

The slope sensitivity test is among the more common ones used for expectations hypotheses. To illustrate this application, this test was used by Mankiw and Miron (1986), among others. They tested equation (26) for the United States using three month and six month data for five intervals during 1890–1979. The null hypothesis ($\beta = 2$) was rejected for all except the earliest period prior to the founding of the Federal Reserve System in 1915. That is, the spread between the short and the long rate was a good indicator of the path of interest rates prior to the commencement of the stabilization operations of the Fed, but not in the periods after it, with the spot rate following a random walk after 1915. Therefore, the authors concluded that a central bank policy of interest rate stabilization would make the spot rate follow a random walk and lead to a rejection of the expectations hypothesis. In general, there seems to be more empirical support for (26) when countries do not pursue interest rate stabilization policies.

20.5.2 Efficient and rational information usage test

Another test of the above joint hypothesis is based on the following restatement of (23):

$$ {}_t\phi^{\mathrm{e}}{}_{t+1} \equiv 2\,{}_tR_{t+1} - {}_{t+1}r^{\mathrm{e}}{}_{t+1} - {}_tr_t = \gamma \tag{27} $$

where ${}_t\phi^{\mathrm{e}}{}_{t+1}$ is the *excess yield* over two periods, which, under the constant liquidity premium version of the expectations hypothesis, equals γ. Assuming REH,

$$ E_t\,{}_t\phi_{t+1} = 2\,{}_tR_{t+1} - E_t\,{}_{t+1}r_{t+1} - {}_tr_t = \gamma \tag{28} $$

The stochastic form of this equation is:

$$ \begin{aligned} {}_t\phi_{t+1} &= 2\,{}_tR_{t+1} - {}_{t+1}r_{t+1} - {}_t r_t + \mu_{t+1} \\ &= \gamma + \mu_{t+1} \end{aligned} \tag{29} $$

where μ_{t+1} is again a random term with $E_t\,(\mu_{t+1}|I_t) = 0$, and I_t is the information available in t. Under the joint hypothesis, the excess yield would not be a function of information known in period t. If a regression of the excess yield on information known in t – such as on prices, output, unemployment, and other variables on which information is commonly available in t – yields significant coefficients for such variables, the joint hypothesis will be rejected by the data. From (29), the regression equation can be formulated as:

$$ {}_t\phi_{t+1} = \alpha + \boldsymbol{b}\boldsymbol{X}_t + \mu_{t+1} \tag{30} $$

where $\alpha = \gamma$, $\boldsymbol{X}_t$ is a vector of commonly known variables in t and $\boldsymbol{b}$ is the corresponding vector of coefficients. Among the $\boldsymbol{X}$ variables would be included the lagged values of the excess yield itself. The maintained hypothesis would be rejected if any of the estimated coefficients in $\boldsymbol{b}$ were significantly different from zero.

Alternatively, in (29), define:

$$ {}_t\phi'_{t+1} = 2\,{}_tR_{t+1} - {}_{t+1}r^{e}{}_{t+1} \tag{31} $$

and specify the regression equation as:

$$ {}_t\phi'_{t+1} = \alpha + \beta\,{}_tr_t + \boldsymbol{bX}_t + \mu_{t+1} \tag{32} $$

where $\beta = 1$. The joint hypothesis would be rejected if the estimated value of β were significantly different from one and/or if any of the b coefficients were significantly different from zero. Jones and Roley (1983) tested (32) for quarterly US Treasury bill data for the period 1970 to 1979. The coefficients of some of the X_t variables were significant, so that the joint hypothesis was rejected.

Many studies using the notion of already available information to test the joint hypothesis for changes in the term structure tend to reject it.[6] A rejection of the joint hypothesis could be due to rejection of the expectations hypothesis, of the assumption of a constant liquidity premium, of the REH, of the proxy used for the expected rate of inflation, or of any combination of them. Therefore, it is not clear whether the expectations theory itself is at fault, since the rejection of the joint hypothesis is sometimes interpreted as a rejection of the assumption of a constant risk premium, sometimes of the REH and sometimes of the proxy used for the expected inflation rate.

The rejection of the assumption of the constancy of the liquidity premium per period implies that this premium can vary over time. This is not implausible for bonds of medium or long maturity, but there is no particular reason to assume that the liquidity premium would vary significantly over periods as short as a week or a few months. Since many of the rejections of the joint hypothesis occur for data using Treasury bill yields only, such rejection may be due to that of the REH or the expectations hypothesis itself. Further, if the liquidity premia are not constant, then the theory needs to specify their determinants, which is difficult to do.

20.6 Random walk hypothesis of the long rates of interest

Start with equation (10). With R and r now redefined as gross rates of interest, (10) becomes:

$$ \begin{aligned} {}_tR_{t+n} &= (1/(n+1))[r_t + {}_tr^{e}{}_{t+1} + {}_tr^{e}{}_{t+2} + \cdots + {}_tr^{e}{}_{t+n}] \\ &\quad + (1/(n+1))\,{}_t\gamma_{t+n}(n;\rho) \qquad n \geq 1 \end{aligned} \tag{33} $$

Lagging (33) by one-period:

$$ \begin{aligned} {}_{t-1}R_{t+n-1} &= (1/(n+1))[r_{t-1} + {}_{t-1}r^{e}{}_t + {}_{t-1}r^{e}{}_{t+1} + \cdots + {}_{t-1}r^{e}{}_{t+n-1}] \\ &\quad + (1/(n+1))\,{}_{t-1}\gamma_{t+n-1}(n;\rho) \end{aligned} \tag{34} $$

6 However, Pesando (1978) reported support for it in Canadian data. This study is discussed in the next section.

Subtract (34) from (33). Applying the REH to the resulting equation gives:

$$
\begin{aligned}
{}_{t}R_{t+n} - {}_{t-1}R_{t+n-1} = (1/(n+1))[(r_t - E_{t-1\,t-1}r_t) + (E_{t\,t}r_{t+1} - E_{t-1\,t-1}r_{t+1})] + \cdots \\
+ (E_{t\,t}r_{t+n-1}) - E_{t-1\,t-1}\,r_{t+n-1})] + (1/(n+1))[(E_{t\,t}\,r_{t+n} - r_{t-1})] \\
+ (1/(n+1))[{}_{t}\gamma_{t+n}(n;\rho) - {}_{t-1}\gamma_{t+n-1}(n,\rho)]
\end{aligned} \tag{35}
$$

Assuming that no new information becomes available in period t – that is, $I_t = I_{t-1}$, where I_t is the information available in t – we have:

$$
E_{t\,t}r_{t+i} - E_{t-1\,t-1}r_{t+i} = \mu_{t+i} \qquad i = 0, 1, \ldots, n-1 \tag{36}
$$

where μ_{t+i} are forecasting random errors with a zero mean and are independently distributed. That is, the revisions to expectations are zero-mean independent random variables.

Further, as $n \to \infty$,

$$
(1/(n+1))[(E_t\, r_{t+n} - r_{t-1})] \to 0 \tag{37}
$$

so that, for large n, this term would be about zero for the usually observed and expected range of values of the interest rate. Hence, if the liquidity premium term on the right-hand side was also a random term or if it equaled zero, $[{}_{t}R_{t+n} - {}_{t-1}R_{t+n-1}]$ would behave randomly. Noting that the last term on the right-hand side of (35) involves the difference between the n-period liquidity premiums in t and $t-1$, an assumption that the liquidity premium is time invariant –that is, does not change with new information – would make this term equal to zero. Alternatively, it can be assumed that, in the absence of any new information, the last term in (35) will also be randomly distributed.

Hence, under the above collection of assumptions, the right hand side of (35) will be a random variable, so that it can be rewritten as:

$$
{}_{t}R_{t+n} = {}_{t-1}R_{t+n-1} + \varepsilon_t \tag{38}
$$

where ε_t is a random error, made up of the relevant set of random errors, with $E(\varepsilon_t | I_{t-1}) = 0$. (38) states that for large values of n the long rates will follow a random walk. This constitutes the random walk hypothesis (RWH) of the long interest rates. Note that it is expected to hold only for large values of n and if no new information becomes available between t and $t-1$.

Since systematic monetary policy followed in t will be anticipated in $t-1$, (38) implies that it cannot affect the change in the long rate between periods $t+n$ and $t+n-1$. Only unanticipated monetary policy – that is, policy shocks which change the value of ε – can shift this difference and shift the yield curve. Hence, the RWH of the long interest rates implies that systematic monetary policy cannot shift the yield curve; only unanticipated monetary policy can do so.

However, since we needed (37) to arrive at (38), note that the RWH does not hold for low values of n. To illustrate the failure of the RWH of long interest rates for low values of n, assume a deterministic system so that $\mu_{t+i} = 0$ and ${}_{t}\gamma_{t+i} = i\gamma$. Further, assume that there is a shift in fundamental factors – with the shift factor β already known in period $t-1$ – such that:

$$
r_{t+1} = \cdots = r_{t+n} = r_t + \beta \tag{39}
$$

where β is the amount of the shift. Given these assumptions, (39) and (35), for $n = 1$, imply that the evolution of the long rate on two-period bonds would be given by:

$$\begin{aligned} {}_tR_{t+1} - {}_{t-1}R_t &= (1/2)(r_t + \beta - r_t) \\ &= (1/2)\beta \end{aligned} \tag{40}$$

We also have:

$${}_tR_{t+2} - {}_{t-1}R_{t+1} = (1/3)\beta \tag{41}$$

and so on to:

$${}_tR_{t+n} - {}_{t-1}R_{t+n-1} = (1/(n+1))\beta \tag{42}$$

where the right-hand side goes to zero only as $n \to \infty$, so that either the RWH must be confined to very large values of n or we must assume that there is no change ($\beta = 0$) in the fundamental or systematic determinants of the long rates.

Pesando (1978) tested the random walk hypothesis – with the assumption of a time-invariant liquidity premium term – in the form of the difference between the rationally expected long yield ($E_{t-1}\ {}_t\mathrm{R}_{t+n}|I_{t-1}$), based on information in $t-1$, and the long yield ${}_{t-1}R_{t+n-1}$. His dependent variable, therefore, was $[(E_{t-1}\,{}_tR_{t+n}|I_{t-1}) - {}_{t-1}R_{t+n-1}]$, which was implied by the joint hypothesis to be random and uncorrelated with information available at the beginning of the period. His regressions of this variable were done against a number of variables such as investment and saving, government deficit, deficit on the current account of the balance of payments,[7] real monetary base and real GNP, and the current change in the monetary base.[8] Pesando's tests on Canadian ten-year bond yields for the periods 1961:1 to 1971:2 and 1961:1 to 1976:4 showed insignificant and/or incorrectly signed coefficients for these variables, so that he could not reject the hypothesis that the current change in the long-term bond yield is a random variable and follows a martingale sequence. Pesando's tests of the models used by Sargent (1969), Echols and Elliot (1976) and Feldstein and Eckstein (1970) for his Canadian data set led to their rejection. These results also rejected the notion that the long yield includes a cyclical term premium determined by these variables, thereby lending support to the hypothesis of a time-invariant premium. Further, his tests rejected the autoregressive procedure for modeling the expected inflation rate. Pesando also rejected the Modigliani and Sutch preferred habitat model since this model requires that the liquidity premium is not time invariant.

The alternative assumption to the above random walk hypothesis is that the changes in long yields depend on economic variables.

7 These variables were also included in the Echols and Elliot (1976) study.

8 These variables are part of the liquidity preference approach and were also included in the Feldstein and Eckstein (1970) study discussed in the preceding chapter.

20.7 Information content of the term structure for the expected rates of inflation

Fisher's relationship between the nominal and real forward interest rates in efficient markets can be specified as:

$$ {}_t r^{\mathrm{f}}{}_{t+i} = {}_t r^{\mathrm{re}}{}_{t+i} + {}_t \pi^{\mathrm{e}}{}_{t+i} \tag{43} $$

where:

$ {}_t r^{\mathrm{f}}{}_{t+i} $ = forward market (nominal) rate of interest in period $t+i$

$ {}_t r^{\mathrm{re}}{}_{t+i} $ = expected real rate of interest in period $t+i$

$ {}_t \pi^{\mathrm{e}}{}_{t+i} $ = expected rate of inflation in period $t+i$.

Restate (43) as:

$$ {}_t \pi^{\mathrm{e}}{}_{t+i} = {}_t r^{\mathrm{f}}{}_{t+i} - {}_t r^{\mathrm{re}}{}_{t+i} \tag{44} $$

Now use the REH to specify the following relationships:

$$ {}_t r^{\mathrm{re}}{}_{t+i} = r^{\mathrm{r}}{}_{t+i} - \eta_{t+i} \tag{45} $$

$$ {}_t \pi^{\mathrm{e}}{}_{t+i} = E_t \pi_{t+i} \tag{46} $$

Equation (44) can now be rewritten as:

$$ E_t \, {}_t \pi_{t+i} = {}_t r^{\mathrm{f}}{}_{t+i} - {}_t r^{\mathrm{r}}{}_{t+i} + \eta_{t+i} \tag{47} $$

which can be restated as:

$$ E_t \, {}_t \pi_{\tau+i} = {}_t r^{\mathrm{f}}{}_{t+i} - E_t \, {}_t r^{\mathrm{r}}{}_{t+i} \tag{48} $$

Equation (48) uses the information on the nominal forward rates, which are known, and the rationally expected value of the *real* interest rates – assuming the latter to be known or already estimated – to derive the expected inflation rates and their term structure over future periods.

Common hypotheses on the real rate of interest

Equations (47) and (48) require data on or estimates of the future real rate $r^{\mathrm{r}}{}_{t+i}$. The range of choices here is usually as follows.

(i) Market data is available on it; e.g. if there is an adequate variety of inflation-indexed bonds in the economy.

(ii) Market data on the future real rate is not available but it can be reasonably assumed to be constant, or that changes in it are very small relative to changes in the inflation rate.

(iii) The assumption of its constancy is not plausible but it is a function of a small number of determinants and this function can be estimated. For example, the loanable funds or the liquidity preference theory, as discussed in Chapter 19, or the more general

IS–LM approach can be used to specify the determinants of this rate. Thus, the IS–LM approach implies that the real money supply and the real fiscal deficit are among the short-run determinants of the real interest rate. The appropriate function for it can be specified and estimated, and the estimated values then substituted in (47) or (48). Among the studies that use this approach to the real rate are those by Sargent (1969), Echols and Elliot (1976), Feldstein and Eckstein (1970) and Pesando (1978). Alternatively, the real interest rate may be assumed to be set by the central bank under a specified Taylor rule.

For an example of (ii), Walsh (2003, p. 498) reports the findings of a 1996 study by Buttiglione, Del Giovane and Tristani. This study examined the impact of monetary policy on long-term interest rates under the assumption that monetary policy does not affect real rates far in the future, so that such future real rates can be taken to be constant. Therefore, the change in long-term interest rates between future periods can be used to estimate the expected inflation rates. Walsh reported the findings of the Buttiglione *et al.* (1996) study as follows: for countries with low average inflation, a contractionary monetary policy raised short rates but lowered forward ones; whereas, for countries with high average inflation, the rise in short rates did not necessarily yield lower forward rates. Walsh points out that the finding for countries with low average inflation is consistent with the hypothesis that, in these countries, the restrictive monetary policy was viewed as a credible policy to lower inflation. Hence, the impact of monetary policy on long interest rates depends not only on the policy pursued but also on the evaluation by the public of its impact on the inflation rate.

An illustration of the empirical results

Mishkin (1990) examined the rates of inflation implied by the yields on one- to five-year bonds in the United States. His data used was monthly from 1953 to 1987. His estimation equation was:

$$ {}_t\pi_{t+i} - {}_t\pi_t = \alpha_{t+i} + \beta_{t+i}({}_tR_{t+i} - {}_tR_t) + \mu_{t+i} \tag{49} $$

where the difference ("inflation spread") in the average annual inflation rate over $t+i$ years was regressed on the spread between the corresponding nominal long rates. Mishkin argued that a rejection of $\beta_{t+i} = 0$ implies that the term structure contains information on the inflation spread, while a rejection of $\beta_{t+i} = 1$ implies that spread in the real rates is not constant over time, in which case the nominal spread in the interest rates does not have any information on the inflation spread, so that nominal spread would provide information on the real spread. His estimates showed that, for the longer maturities, the spread in nominal interest rates contains substantial information about the inflation spread but little about the real interest rate spread. Contrary to the findings of many other studies, the converse was found for short maturities of six months or less. For these, the term structure did not contain any information on the future change in inflation but did imply a significant amount about the term structure of *real* rates of interest.

Barr and Campbell (1997) provide an example of the derivation of expected future increases in inflation from data on long rates. For the UK, they use the yield on indexed (i.e. with the nominal interest rising to compensate for the inflation rate) and nominal (un-indexed) government bonds to estimate the expected inflation rates. Their finding is

that about 80 percent of the changes in the long-term nominal interest rates reflect expected long-term inflation.

Note that the spread between the interest rates on short-term and long-term bonds and those on commercial paper and Treasury bills can have predictive power, not only for expected inflation but also for future output changes, as reported, for example, by Stock and Watson (1989) and Friedman and Kuttner (1992).

Conclusions

It seems clear that in well-established financial markets, the markets in the individual assets are closely interrelated. For such markets, while the risk premium may vary, possibly in a narrow range, the likelihood that the markets for government bonds of different maturities are very significantly segmented seems to be small. Hence, the expectations hypothesis of the term structure of interest rates is generally the maintained hypothesis for closely integrated markets. This assumes that there is very extensive substitution among the bonds of different maturities (with other characteristics such as risk held constant), the markets are efficient and the transactions costs are not significant, at least for the dominant economic agents in the financial markets.

Substitution among the securities of different maturities is likely to be limited where the markets are thin or not well developed, as in the less-developed economies. The degree of substitution existing in the markets is, therefore, primarily a function of the stage of economic and financial development of the country.

Empirical tests of the expectations theory are usually of a combined or joint hypothesis including the expectations theory, the assumption of a constant liquidity premium and the rational expectations hypothesis. These tests usually use reduced-form equations for the yields on different maturities or their spread. Most such studies prove to be unfavorable to the joint hypothesis, with some supporting it, but it is not clear whether the expectations hypothesis itself is at fault. The problem could lie with the assumption of a constant liquidity premium or the assumption of rational expectations.

Several of the tests of the joint hypothesis use Treasury bill rate data, which is unlikely to incorporate much variation in the liquidity premia. Also, in general, tests on the rational expectations assumption do not provide unambiguous support for it.

Tests of the alternative hypothesis – that the markets are compartmentalized by the term to maturity – usually use structural systems of demand and supply equations. If this segmented markets or preferred habitat hypothesis were correct, the debt management policies of the government should be able to shift the yield curve. However, empirical studies do not support the idea of a significant shift in the yield curve brought about by changes in debt management policies.

Therefore, for the developed financial markets, while the more rigorous and detailed tests of the joint hypothesis often reject it, there is not sufficient evidence to individually support findings of significant variations in the liquidity premium, or to provide sufficient support for the alternative hypotheses of the segmented markets or preferred habitats. However, there may be some small possibility of these effects, which, along with small deviations in expectations from the rational ones, could be enough to reject the joint hypothesis. From the opposite perspective, there is a great deal of evidence that most of the variations in the long rates can be explained by variations in the expected future short rates. Since the latter are closely related to the expected rate of inflation for the relevant period, the long rates are often used as good (even though not perfect) predictors of the expected inflation rates.

Summary of critical conclusions

- ❖ While the yield curve can have any shape, its shape at zero current and expected inflation rates is usually upward sloping.
- ❖ Long rates are geometric averages of the current and forward short rates.
- ❖ The main theory of the yield curve is the expectations theory. It assumes that the bond markets are efficient and have zero transactions costs.
- ❖ Liquidity preference in the context of the term structure of interest rates asserts that bonds of longer maturity incorporate a premium over those of shorter maturity to compensate for their lesser liquidity.
- ❖ The segmented and preferred habitat hypotheses assume that there exist significant preferences among the borrowers and lenders for specific maturities, rather than indifference or a fairly weak preference for shorter over longer maturities.
- ❖ The expectations hypothesis and rational expectations imply, under certain other assumptions, the random walk hypothesis for changes in the long rates of interest.
- ❖ The yield curve can be used to estimate the expected rates of inflation over future periods.

Review and discussion questions

1. The yield curve shows the relationship between the yields of high-grade securities that differ only in the term to maturity. Sometimes the curve rises, sometimes it falls and sometimes it is flat. Present the main reasons for these different shapes.
2. "On the basis of recent empirical studies, the expectations hypothesis with efficient markets and rational expectations does not seem to explain the term structure of interest rates." Discuss. Present the findings of at least one such study and discuss the potential reasons for this failure.
3. Can the central bank change the shape of the yield curve through changes in (a) the term structure of government bonds and (b) variations over time in monetary aggregates? Discuss.
4. For your country, what is the current shape of the yield curve? Explain this shape.
5. Assuming that the expectations theory of the yield curve is correct, derive from your data on the yield curve the expected future spot rates for the next twelve months. If your economy has forward markets for this period, compare the forward short rates with your derived expected future spot rates, and explain the reasons for any differences.
6. Use your data on the yield curve to derive the expected rates of inflation monthly over the next twelve months and annually over the next five years. What assumptions were needed for this derivation? If the actual inflation rates turn out to differ from your computed ones, how would you explain the differences?
7. How does the inability to establish the yield curve for the real rate of interest affect the derivation of the future expected rates of inflation from the nominal interest rates? Does the existence of a term structure of liquidity premiums pose corresponding problems?
8. For the derivation of the expected future inflation rates from the yield curve, is it really valid to assume (a) a constant real rate of interest, (b) a constant liquidity premium per period? Cite some studies that have done so, and report on their findings.

What are the determinants of the real rate interest and the liquidity premium? Does the expected stance of monetary or fiscal policy affect these variables?

9. "The theory of portfolio selection has little relevance for explaining the demand for money. Its real relevance is to the theory of bond and equity returns and prices." Discuss. Show how it can be used to explain these returns and prices.
10. For two-period bonds, show that the expectations hypothesis approximately implies that:

$$R_2 = \tfrac{1}{2}r_1 + \tfrac{1}{2}{}_1r^e{}_2$$

where R and r are the logs of the relevant gross rates.

It is sometimes claimed that long rates overreact to current short rates because financial markets are "myopic." Define what you understand by myopia in this context. How would you modify the above equation to allow for such myopia? How would you test the resulting hypothesis?

References

Barr, D.G., and Campbell, J.Y. "Inflation, real interest rates, and the bond market: a study of UK nominal and index-linked government bond prices." *Journal of Monetary Economics,* 39, 1997, pp. 361–83.

Buttiglione, L., Del Giovane, P. and Tristani, O. "Monetary policy actions and the terms structure of interest rates: a cross-section analysis," Paper presented at the Banca d'Italia, IGIER and Centro Paolo Baffi workshop, *Monetary Policy and the Term Structure of Interest Rates*, Universita Bocconi, Milan, June 1996.

Culbertson, J.M. "The term structure of interest rates." *Quarterly Journal of Economics*, 71, 1957, pp. 485–517.

Echols, M.E., and Elliot, J.W. "Rational expectations in a disequilibrium model of the term structure." *American Economic Review*, 66, 1976, pp. 28–44.

Feldstein, M., and Eckstein, O. "The fundamental determinants of the interest rate." *Review of Economics and Statistics*, 52, 1970, pp. 363–75.

Fisher, I. *The Theory of Interest*. New York: Macmillan, 1930.

Friedman, B.M. "Financial flow variables and the short-run determination of long-term interest rates." *Journal of Political Economy*, 85, 1977, pp. 661–90.

Friedman, B.M., and Kuttner, K.N. "Money, income, prices and interest rates." *American Economic Review*, 82, 1992, pp. 472–92.

Hicks, J.R. *Value and Capital*. Oxford: Oxford University Press, 1946.

Jagannathan, R., and McGrattan, E.R. "The CAPM debate." *Federal Reserve Bank of Minneapolis Quarterly Review*, 1995, pp. 2–17.

Jones, D.S., and Roley, V.V. "Rational expectations and the expectations model of the term structure: a test using weekly data." *Journal of Monetary Economics*, 12, 1983, pp. 453–65.

Mankiw, N.G., and Miron, J.A. "The changing behavior of the term structure of interest rates." *Quarterly Journal of Economics,* 60, 1986, pp. 211–28.

Meiselman, D. *The Term Structure of Interest Rates*. Englewood Cliffs, NJ: Prentice-Hall, 1962.

Mishkin, F.S. "What does the term structure tell us about future inflation?" *Quarterly Journal of Economics,* 105, 1990, pp. 815–28.

Modigliani, F., and Sutch, R. "Innovations in interest rate policy." *American Economic Review Papers and Proceedings*, 56, 1966, pp. 178–97.

Modigliani, F., and Sutch, R. "Debt management and the term structure of interest rates: an empirical analysis of recent experience." *Journal of Political Economy*, 75, 1967, pp. 568–89.

Pesando, J.E. "On the efficiency of the bond market: some Canadian evidence." *Journal of Political Economy*, 86, 1978, pp. 1057–76.

Sargent, T.J. "Commodity price expectations and the interest rate." *Quarterly Journal of Economics*, 83, 1969, pp. 127–40.

Sharpe, W.F. "Capital asset prices: a theory of market equilibrium under conditions of risk." *Journal of Finance*, 19, 1964, pp. 425–42.

Stock, J.H., and Watson, M.W. "Interpreting the evidence on money–income causality." *Journal of Econometrics*, 40, 1989, pp. 161–81.

Walsh, C.E. *Monetary Theory and Policy*. Cambridge, MA: MIT Press, 2003.

Part VII

Overlapping generations models of money

21 The benchmark overlapping generations model of fiat money

Overlapping generations (OLG) models of money have been proposed by some economists as an alternative to the money in the utility function (MIUF) and money in the production function (MIPF) models presented in Chapter 3. However, other economists do not consider the OLG models of money in their standard form to be valid or useful for modeling the actual role of money in the economy.

This is the first of three chapters using the OLG framework to model the role of money in the economy. This chapter and the next present the pure versions of the OLG models, while the third chapter attempts to integrate the concepts of cash-in-advance, MIUF and MIPF into the OLG format.

Key concepts introduced in this chapter

- Fiat money
- Overlapping generations
- Intrinsically useless money
- Inconvertible fiat money
- Market fundamentals
- Sunspots
- Bubbles
- Bootstrap paths

This chapter and the next two present overlapping generations (OLG) models of money. These models have been proposed as an alternative to the MIUF (money in the utility function) and MIPF (money in the production function) models presented in Chapter 3 and used in the earlier chapters of this book. Since there are now two competing sets of models, we need to establish the criteria for selecting between them. Given the mandate of economics as a science to explain real-world observations, this selection should be based on the degree to which the competing models explain the stylized empirical findings on money in the modern economy.

The models presented so far in this book to explain holdings of money by an individual and the economy have been one-period (timeless) ones. This chapter and the next two separate the individual's lifetime into two (or more) "lifestages," with two (or more) generations alive in the economy simultaneously and with trading between the generations. Such models are called *overlapping generations models*.

The concept of overlapping generations for the analysis of money in the economy was introduced by Samuelson (1958),[1] with major extensions by Wallace (1980, 1981) and Sargent (1987), among others: Champ and Freeman (1994) present a textbook application of this approach. The standard version of the OLG model assumes that the individuals in the economy live for two periods only – or for two lifestages, "young" and "old," with each lifestage lasting one period – and that in each period the economy has two generations of individuals. One of these is the old generation of individuals who were born in the preceding period and the other is the young generation born at the beginning of the current period. The old of one generation and the young of the next one overlap in every period, so that the name given to the models using this framework is the overlapping generations models.

The OLG framework is a substitute for a timeless or an infinite one, with the representative agent having an infinite horizon. It does not by itself provide a model, but has to be combined with other assumptions in order to yield a meaningful model. This chapter presents a benchmark model with the OLG framework and certain assumptions about money, endowments, etc. This model is based on the work of Wallace (1980, 1981) and is referred to in this and the next two chapters as the basic, standard or benchmark OLG model with money. The basic form of this model and its implications are derived in this chapter.

Section 21.1 specifies the stylized facts, against which the validity of the implications of all models with money needs to be judged, about money in the modern economy. Section 21.2 presents the common themes regarding money in the usual OLG models with money. Sections 21.3 to 21.6 present the basic OLG model with money. Sections 21.7 and 21.8 show the inefficiency of monetary expansion in such models, even when such expansion is needed to ensure a stable price level. Sections 21.9 and 21.10 derive money demand when there is a positive rate of time preference and when there are several fiat monies. Section 21.11 deals with sunspots, bubbles and market fundamentals.

21.1 Stylized empirical facts about money in the modern economy

For assessing the validity of OLG models discussed in this chapter and in Chapters 22 and 23, the stylized facts related to money are:

1 Money is more liquid than other financial assets (bonds) and commodities in the sense that it has lower transactions costs in making payments than the latter.
2 Money consists of inside money (liabilities of the private sector) and outside money (the monetary base). Only the latter is fiat money. Therefore, fiat money and private substitutes to fiat money coexist and circulate simultaneously in the economy.
3 Every economic agent holds money in every period, irrespective of whether saving during any of the periods is positive, negative or zero. If we were to divide the individual's

1 Samuelson presented a three-period (along with a simplified two-period) version of the OLG model, with his focus being on the rate of interest. In the three-period model, individuals worked and earned income in the first two periods and were retired in the third period. His problem was to allow positive consumption in the third period, when there was zero income and the commodities could not be stored over time. Samuelson showed how the economy could achieve this pattern of consumption through the social contrivance of fiat money, with the latter bringing about a "biological rate of interest" equal to the population growth rate.

lifetime into three stages, pre-job,[2] working and retired, money is held in each of the three life stages.

4 The main scale determinant of the demand for money is current income or wealth (or its proxy, permanent income). In particular, money demand by households is closely related to their consumption expenditures. It is not closely related to their current saving.
5 The price level in the economy is determined by the money supply relative to national income/output, rather than by the money supply relative to current saving.
6 The velocity of money is normally greater than one and fluctuates over time.
7 A monetary economy does not function in a stationary state. There is continual change in output, prices, expected prices, etc.
8 Money has a positive value even if there is inflation, which erodes its value. In particular, money has a positive value even in hyperinflations, in which its value decreases rapidly.
9 The demand for money is positive even though the economy has riskless bonds with a positive coupon payment (e.g. Treasury bills), and even though money may not pay interest.

The MIUF and MIPF models, with a timeless or infinite horizon for the representative agent, yield implications consistent with these stylized empirical facts on money in the economy. This chapter derives the implications of the benchmark OLG model of money to evaluate its realism by comparison with the above stylized facts.

21.2 Common themes about money in OLG models

Fiat money: intrinsically useless and inconvertible

The OLG models with money attribute to fiat money two basic characteristics: (a) it is *intrinsically useless* – that is, it cannot be directly used in production or consumption[3] – and (b) it is *inconvertible*[4] – that is, the *issuer* does not give a commitment to redeem it into commodities. Further, it is assumed that fiat money is costless to produce, to store and to transfer, and does not pay interest to the holder. The common example of fiat money is banknotes issued by the central bank. In most modern economies, such notes do not carry a commitment by the central bank to redeem them into gold coins or gold bullion or into any other commodity. Given the emphasis on the nature of fiat money as being

2 The pre-job and retirement stages usually do not have labor income. The retirement stage usually has negative saving so that assets are run down. The pre-job stage does not usually have any asset income, and its expenditures are often financed from handouts from parents.

3 The "intrinsically useless" concept is an inherently flawed one for economic analysis since it is difficult, if not impossible, to define what is "directly used in consumption or production" and what is not. For example, are diamonds, cigarettes and drugs intrinsically useful? Is their value determined by their intrinsic usefulness? What *does* confer "usefulness" at the margin of demand?

Note also that bonds would also be intrinsically useless since they are not directly used in consumption or production.

4 The validity of this concept for the individual is also of questionable merit. The individual does not care and usually does not even know whether the central bank – the issuer of fiat money – will ever redeem its notes into commodities, as long as others from whom he wants to buy commodities will do so. That is, the individual only requires convertibility of fiat money against commodities in the private markets.

Note that consols are also inconvertible since the central bank, which issues such bonds, does not give a commitment to ever redeem them.

intrinsically useless, the standard OLG models assume that money does not appear in the utility and production functions. For brevity, we will designate the term "OLG model" to implicitly include this feature.

A positive value for fiat money

One of the necessary conditions for fiat money, or any other good, to have a positive market value is that its supply and that of its close substitutes should be limited and strictly regulated, or that its production involves some cost. OLG models with money therefore assume that the supply of fiat money, though costless to produce, is limited, which requires, in turn, that it cannot be easily and costlessly counterfeited and the counterfeit notes put into circulation to any significant extent,[5] and that the private sector cannot costlessly create close substitutes or near-monies and put them into circulation.[6]

In addition, given the intrinsic uselessness and inconvertibility of fiat money, it is a characteristic of all monetary models – whether OLG, MIUF or others – that fiat money will have value in exchange only if others are willing to accept it in exchange for commodities. Given the intergenerational emphasis of the OLG models, these models narrow this condition to imply that fiat money will have value in exchange in any *given* period only if it is expected that the individuals in the *following* period will be willing to accept the fiat money in exchange for commodities. This requires that if the economy is currently to possess a positive value of money, it must be expected to continue indefinitely into the future, so that the assumed model must *not* have a finite horizon.[7]

Furthermore, while all monetary models imply that expectations about the future value of the money are one of the determinants of its current value, the OLG models imply the stronger condition that if the next period has a zero value for money, its current value must also be zero. By extension, if fiat money is expected to be valueless in any period T in the future, it will also have a zero value in the current period and in all periods up to T. As against this conclusion of the OLG models, MIUF models do not require an infinite horizon, nor do they imply that money cannot have a positive value, determined by its usefulness in current exchanges, even if its value in some future period is zero.

Fiat money as the medium for holding savings

In OLG models with money, a necessary condition for fiat money to have a positive value in the current period is that there currently exist individuals who want to use it to carry purchasing power from the present to the future. This requires that they want to consume

5 Alternatively, the cost of counterfeiting a note must exceed the value of the note in exchange.

6 While modern economies allow the state to monopolize the creation of currency and make it illegal to counterfeit it, the creation of near-monies is permitted and private financial intermediaries create a variety of substitutes for money. One of these close substitutes is checking deposits. Other creations of the private sector allow an increase in the velocity of circulation of money. Among these have been the creation of credit cards, debit cards, automatic teller machines, electronic transfer of deposits, etc. These activities of the private sector limit the applicability of the standard OLG models.

7 With a finite horizon, money would be worthless at the end of the terminal period. To avoid a 100 percent capital loss, individuals in the terminal period will not want to obtain and hold money, thereby rendering it worthless for individuals who would obtain it one period earlier and hold it to the terminal period, and so on for still earlier periods, thus making the demand for money zero in all periods.

more in the future (when they are old) than their future receipts of commodities and that, to do this, they have positive saving – i.e. an excess of commodities over their consumption – in their young lifestage. It also implies that money is an attractive vehicle for saving. In order to ensure the former requirement, the OLG models assume that the receipt of commodities in the first (or young) lifestage exceeds their optimal consumption, so that the individuals must have positive saving in the first lifestage to carry over for consumption in the second (or old) lifestage.

Saving out of the current endowment of commodities is initially in the form of commodities. If the carryover of commodities were costless, or consumption in the future could be financed at lower cost through some contrivance associated with commodities rather than with the use of fiat money, fiat money would not be used and again would not have positive value.[8] Therefore, the usual OLG models assume that fiat money does not have storage or transfer costs while commodities do so. In the limiting case, commodities are assumed to be perishable.

The essential assumptions and implications of the OLG models with fiat money

The assumptions of the standard OLG models with money are:

1 Defining bonds as *interest-bearing* financial assets that can be used to convey purchasing power from the present to the future, there are no bonds in the model.[9]
2 Fiat money is preferable to commodities – and any other assets – as the medium for carrying forward saving to the following period.
3 There is net (positive) saving in the first lifestage.
4 Future periods will not renounce the use of fiat money or pursue policies such that fiat money will become valueless.
5 The OLG model's economy has an infinite horizon, even though the individuals in it have a finite (two-period) horizon.

Given these assumptions, the OLG models of fiat money explore the value of money for various growth rates of money versus commodities, growth of population, open market operations, etc.

Among the attractive features claimed for OLG models is that, along certain paths, they establish a positive value for an intrinsically worthless fiat money[10] which is not required by law to be convertible into commodities, and that time and the distinctiveness of the earning pattern over a lifetime are incorporated in an "essential" manner.[11] Further, they allow for economic agents who are identical at birth – thus permitting the study of stationary

8 Therefore, the return on commodities, net of any premium for risk and depreciation from storage, must be less than on money. This is definitely the case if the commodity is assumed to be perishable, in which case the return on commodities is −100 percent.

9 In fact, fiat money in the OLG models is really a zero-interest bond. It is not money in the usual sense of being a medium of payments among individuals of the same generation since they are assumed to be identical.

10 We have already argued that the concept of intrinsic usefulness is of doubtful use in economics. Further, OLG models are not unique in establishing a positive value of intrinsically worthless goods. All theories (including the MIUF and MIPF) with fiat money do so, as do those that determine the prices of diamonds and cigarettes, etc.

11 None of these properties are unique to the OLG models: the MIUF and other monetary models can be also formulated to possess them.

states – while allowing for a degree of heterogeneity among the economic agents alive at any time in the economy, and also allow – indeed require – the economy to continue indefinitely into the future.

As pointed out already, OLG models with money generate a zero value of money in the current period if the value of money is expected to be zero in some future period. This is a characteristic of *bootstrap* or *bubble paths,* which are paths along which the values of the variables depend upon expected values, even if arbitrary ones, in the future and change if the latter change. The numerous equilibria of this kind are among the *tenuous* kind, meaning by this that they are not based on the fundamentals[12] of the system. However, the usual focus of OLG models is not on such bootstrap or bubble paths. Rather, their implications are normally analyzed only for the stationary states of the economy, with expectations assumed to be identical with the stationary values or with those implied by the rational expectations hypothesis (REH).[13]

21.3 The basic OLG model

In the standard version of the OLG framework, individuals live for two periods – that is, go through two lifestages – only. They are often labeled "young" in their first lifestage and "old" in their second lifestage. This book uses the superscripts y and o to indicate the individual's respective lifestages.

For the economy, the periods are $t + i$, $i = 0, 1, 2, \ldots$. Period t is the initial period of the analysis and its old generation is called the "initial old," whose members were born in period $t - 1$. Generations born in periods $0, 1, 2, \ldots$, will be called the "future generations" and its members will be referred to merely as "individuals." The OLG model starts by endowing the initial old with the initial stock of money. Further, for the basic OLG model of this chapter, it is assumed that any increase in the money stock in any period is gratuitously given as a lump-sum transfer to the old in that period. The next chapter deviates from this assumption to examine the case where the seigniorage from money creation is used to buy up commodities that are then destroyed, resulting in a net decrease in the commodities left for consumption in the economy.

The number of individuals born in period t is N_t. In the early parts of the analysis of this chapter, this number is assumed to be constant at N over time. Under this assumption, in each period t, the population of $2N$ individuals consists of N young individuals and N old individuals.

Each individual is assumed to be given a commodity endowment of w^{y} in the young lifestage and w^{o} in the old lifestage. w^{y} and w^{o} are in units of the single consumption good, assumed in the basic model to be non-storable (perishable).[14] Some of the versions of the OLG models assume that w^{o} is zero, but such an assumption is not essential to the OLG framework. However, if fiat money is to have value, it is essential to assume that the optimal level of consumption in old age will exceed w^{o}.[15] This is usually guaranteed by an assumption that

12 The fundamentals of the economy are preferences, production technology and population.

13 Hence, the determination and properties of bubbles are not analyzed in such models.

14 This assumption is relaxed in many versions of the OLG framework. It is maintained throughout this chapter, but not when presenting the Modigliani–Miller theorem on open market operations in OLG models in the next chapter.

15 An alternative – and common – set of assumptions in the literature is that: the individual supplies an amount n^s of labor in the young lifestage and none in the old one; consumption occurs only in the old lifestage

consumption will be the same in each lifestage and that $w^{o} < w^{y}$, so that the individual must save while young to provide for extra consumption in the second period.[16] This assumption will be implicit throughout this chapter.

21.3.1 Microeconomic behavior: the individual's saving and money demand

Intertemporal budget constraint of the young

In the young lifestage, the representative individual can either consume c^{y} or hold money m out of his endowments of commodities. His budget constraint for the first/young lifestage is:

$$p_t c^{y}{}_t + m^{y}{}_t = p_t w^{y}{}_t \quad c^{y}{}_t < w^{y}{}_t \tag{1}$$[17]

At the beginning of period $t + 1$, the individual has the carryover money balances of m_t (which do not pay interest) and receives gratuitously the (real) endowment of commodities $w^{o}{}_{t+1}$, so that his second/old lifestage constraint is:

$$p_{t+1} c^{o}{}_{t+1} = p_{t+1} w^{o}{}_{t+1} + m^{o}{}_{t+1} \tag{2}$$[18]

where the money balances purchased when young, $m^{y}{}_t$, become the inheritance of the old as $m^{o}{}_{t+1}$, so that $m^{y}{}_t = m^{o}{}_{t+1} = m_t$. Note that there is no explicit interest rate in this model since the commodity is perishable and there are no interest (or coupon) paying assets in the model. The only asset is money, which does not pay interest, so that the interest rate does not enter (2). Note also that the individuals are assumed to have perfect foresight over the future values of the variables. From (2),

$$m^{o}{}_t = p_{t+1} c^{o}{}_{t+1} - p_{t+1} w^{o}{}_{t+1} \tag{2'}$$

Noting that $m^{o}{}_{t+1} = m^{y}{}_t$, substitution of (2′) in (1) gives the individual's lifetime budget constraint as:

$$p_t c^{y}{}_t + p_{t+1} c^{o}{}_{t+1} = p_t w^{y}{}_t + p_{t+1} w^{o}{}_{t+1} \tag{3}$$

Define the individual's real lifetime wealth W_t as:

$$W_t = w^{y}{}_t + (p_{t+1}/p_t) w^{o}{}_{t+1}$$

and equals the amount c, the intertemporal utility function is $[u(c) - v(n)]$, and the production function has constant returns to scale, with one unit of labor producing one unit of the consumer good. The qualitative implications of these assumptions are similar to those where labor is supplied and consumption occurs in both periods, provided there is net saving in the first lifestage. The latter correspond to our assumptions in the text.

16 Alternatively, the endowments in the two periods can be specified in terms of the ability to work and a corresponding job. In such an interpretation, the first lifestage would be interpreted as the individual's working life, with income from work, and the second as the retired stage, where there is no labor income but there could be some exogenous income from social security and other public sources.

17 We have specified this constraint as an equality for the rational young individual since he would either consume all his endowments or convert any saving into money m for possible use in the future.

18 Since the individual derives no utility from unspent money balances or unconsumed commodities left over at the end of $t + 1$, utility maximization implies that the old lifestage constraint is an equality.

The symbols used so far and their definitions are:

c^{y}_t	consumption of the young in period t
c^{o}_t	consumption of the old in period t
p_t	price of goods in period t
w^{y}_t	exogenous real income of the young in period t
w^{o}_t	exogenous real income of the old in period t
W_t	lifetime wealth in period t
N_t	number of persons born in period t
$N_{t-1} + N_t$	population in period t
m^{y}_t	*per capita* demand for *nominal* balances by the young in period t
m^{o}_t	money endowment of each old individual in period t
M_t	total amount of fiat money in period $t(= N^{o}_t m^{o}_t)$.

Note that this chapter, as well as Chapters 22 and 23, defines m as the *nominal* (not real) *per capita* balances, contrary to our usage in the rest of this book.

Since $c^{y}_t < w^{y}_t$ by assumption, the young want to transfer commodities to themselves in the future but the non-storable commodity assumption of the model prevents them doing so directly – as it were, through barter (via storage) between themselves when young and when old. Further, the auctioneer and other costless clearing mechanisms of the general Walrasian equilibrium models are excluded from the OLG models. So are state-enforced compulsory exchanges between generations, as through a government pension or social security system. Similarly excluded are private intergenerational mechanisms for transfers of commodities between generations through a private pension plan or an extended family system. The OLG models only allow the transfer of commodities over generations through trade, with the intermediation of money.

The general run of monetary models – and our analyses in Chapters 1 to 20 – allow the use of bonds, which yield a higher positive return than fiat money and act as the intermediary instrument for exchanges of commodities over generations. However, the basic OLG model assumes that there are no other assets such as bonds that can act as a store of value with a rate of return higher than that on fiat money. Consequently, there is no private (or public) borrowing or lending in the basic model[19] and therefore no bonds and no explicit interest rate. The next chapter will present an OLG model with money and bonds.

Utility maximization by the young

The individual born in period t has an intertemporal utility function:

$$U(c^{y}_t, c^{o}_{t+1})$$

where $U(.)$ is assumed to be an ordinal utility function with continuous first- and second-order partial derivatives. The young maximize this intertemporal/lifetime utility function subject to the lifetime budget constraint (3). That is, the young's optimization problem is:

$$\text{Maximize} \qquad U(c^{y}_t, c^{o}_{t+1}) \tag{4}$$

$$\text{subject to:} \qquad p_t c^{y}_t + p_{t+1} c^{o}_{t+1} = p_t w^{y}_t + p_{t+1} w^{o}_{t+1} \tag{3}$$

19 In the two-lifestages models, private but non-negotiable borrowing (in IOUs or bonds) is excluded since any given borrower and lender can meet in only one period and never again, so that they can meet to arrange a loan but can not meet to settle its repayment.

implying the optimal consumption amounts c^{y}_{t}, c^{o}_{t+1} as:

$$c^{y}_{t} = c^{y}_{t}(p_{t+1}/p_t, w^{y}_{t}, w^{o}_{t+1}) \tag{5}$$

$$c^{o}_{t+1} = c^{o}_{t+1}(p_{t+1}/p_t, w^{y}_{t}, w^{o}_{t+1}) \tag{6}$$

By assumption, with $w^{y}_{t} > w^{o}_{t+1}$,

$$c^{y}_{t} < w^{y}_{t}$$

$$c^{o}_{t+1} > w^{o}_{t+1}$$

The net dissaving in the old lifestage is accomplished by spending the money balances carried over from the young lifestage. Optimal saving s^{y}_{t} in period t is given by:

$$\begin{aligned} s^{y}_{t} &= w^{y}_{t} - c^{y}_{t} \\ &= s^{y}_{t}(p_{t+1}/p_t, w^{y}_{t}, w^{o}_{t+1}) \end{aligned} \tag{7}$$

and the demand for money, identical with that for nominal saving, is:

$$m^{y}_{t} = p_t s^{y}_{t} = p_t s^{y}_{t}(p_{t+1}/p_t, w^{y}_{t}, w^{o}_{t+1}) \tag{8}$$

Intuitively, in period t, the young individual receives more of the consumption good than he wants to consume but cannot store the excess since the consumption good is perishable. He sells it to the initial old for fiat money, provided that he expects to be able to exchange his fiat money holdings for the consumption good in period $t+1$.

Utility maximization by the initial old

From the perspective of the initial old in the initial period t, they receive some of the consumption good. Further, while they received fiat money, its utility in consumption is zero so that they are willing to exchange it for some amount of the consumption good. Formally, the utility function and budget constraint, respectively, of the initial old are:

$$U^{o}_{t} = U(c^{o}_{t}) \tag{9′}$$

$$p_t c^{o}_{t} = p_t w^{o}_{t} + m^{o}_{t} \tag{9″}$$

Each member of the initial old maximizes his utility by maximizing c^{o}_{t}, which implies that he will try to trade m^{o}_{t} for the maximum amount that he can get of the consumption good.

21.3.2 Macroeconomic analysis: the price level and the value of money

There are only two goods, the commodity and money, in this OLG model, so that the macroeconomic analysis has to take account of only the markets for money and the commodity. Further, by Walras's law, equilibrium in one of these markets ensures equilibrium in the other one, so that we need to present the analysis of one market only. We choose to focus explicitly on the money market for further analysis.

The market for money and price level determination

For the economy in period t, the aggregate demand for nominal balances $M^{d}{}_{t}$ equals the nominal value of the commodities the young want to sell, so that it is given by:

$$M^{d}{}_{t} = N_t[p_t(w^{y}{}_{t} - c^{y}{}_{t})] \tag{10}$$

The money supply M_t in the economy is given by the money balances held by the old (born in $t-1$ with their number as N_{t-1}). The old want to trade it for commodities. This amount equals:

$$M_t = N_{t-1}[m^{o}{}_{t}] \tag{11}$$

so that money market clearance, with money demand equal to money supply, implies that:

$$N_t[p_t(w^{y}{}_{t} - c^{y}{}_{t})] = M_t \tag{12}$$

Hence:

$$p_t = M_t/[N_t(w^{y}{}_{t} - c^{y}{}_{t})] \tag{13}$$

From (5), $c^{y}{}_{t}$ on the right side of (12) depends on p_{t+1}/p_t, $w^{y}{}_{t}$ and $w^{o}{}_{t+1}$, so that:

$$p_t = M_t/[N_t(w^{y}{}_{t} - c^{y}{}_{t}(p_{t+1}/p_t, w^{y}{}_{t}, w^{o}{}_{t+1}))] \tag{13$'$}$$

Hence, *ceteris paribus*, the price level p_t varies proportionately with the money supply M_t, which is a quantity theory result. Further, note that p_t depends on the intertemporal price ratio p_{t+1}/p_t.

From (13), the value v_t per unit of money, which is equal to $1/p_t$, is given by:

$$\begin{aligned} v_t &= [(w^{y}{}_{t} - c^{y}{}_{t})]N_t/(N_{t-1}m^{o}{}_{t}) \\ &= [(w^{y}{}_{t} - c^{y}{}_{t})N_t]/M_t \end{aligned} \tag{14}$$

where $c^{y}{}_{t} = c^{y}{}_{t}(p_{t+1}/p_t, w^{y}{}_{t}, w^{o}{}_{t+1})$. Hence, the value of money is positive and changes inversely with the money supply. It also varies proportionately with aggregate saving $[(w^{y}{}_{t} - c^{y}{}_{t})N_t]$.

Note that since p_t and v_t are functions of $c^{y}{}_{t}$, which is specified by (5), they are also functions of the future expected price level p_{t+1}. Since the latter is a function of p_{t+2}, and so on, p_t and v_t are functions of the expected prices of the commodity in future periods. The implications of this for price bubbles and the indeterminacy of prices will be analyzed later.

The commodity market

Note that since there are only two goods – the commodity and money – in the model under review, by Walras's law, money market clearance also implies commodity market clearance. Hence, the price level given by (13) also clears the commodity market, so that the derivations of the demand and supply functions in the commodity market are not needed.

21.3.3 The stationary state

Assume an economy with a stationary population, endowments of commodities[20] and money supply. For such an economy, we want to find the path along which the values of all the endogenous variables, such as the price level, consumption, etc., do not change over time. Designate this path as the stationary state. This model has two stationary states: (i) a stationary state with trade (between the young and old generations), and (ii) a stationary state without trade. We start with the analysis of the stationary state with trade. For the following equations, note that, for the stationary state, the subscript t on the variables can be omitted.

Price level in the stationary state with trade between money and the commodity

For the stationary state with trade between money and the commodity, (13) becomes:

$$N[p(w^{y} - c^{y})] = M \tag{15}$$

Hence, the stationary state price level is:

$$p = M/[N(w^{y} - c^{y})] \tag{16}$$

Therefore, since all the variables on the right-hand side have been assumed to be stationary, the price level p is also stationary.

For the stationary state, (16) yields the value v of money as:

$$v = 1/p = [N(w^{y} - c^{y})]/M \quad v > 0 \tag{17}$$

where $c^{y}{}_{t} = c^{y}{}_{t}(p_{t+1}/p_t,\ w^{y}{}_{t},\ w^{o}{}_{t+1})$. The stationary state value of v is positive since $(w^{y} - c^{y}) > 0$.

Equation (13) specifies the *gross* real rate of return on money,[21] r_t, as:

$$r_t = v_{t+1}/v_t = p_t/p_{t+1} \tag{18}$$

so that, in the stationary state:

$$r_t = r = 1$$

Hence, the *net* rate of return – that is, excluding the return of the principal amount lent – on real balances will be zero.

Stationary state without trade between money and the commodity

One of the stationary states is without trade between money and the commodity. Such a state is known as the *autarchic* one. In this state, there is no trade between the young and the old through the contrivance of money, so that the young consume all of their endowments while

20 In the stationary state, the endowments of commodities are identical across generations, but are not identical over the two lifestages of any generation. In fact, it has been assumed that $w^{y}{}_{t} < w^{o}{}_{t+1}$.

21 The gross rate of return or interest over a period specifies the total amount (principal plus interest) that will be received at the end of the period from an investment of $1 at the beginning of the period.

the consumption of the old is limited to their endowments. Money and prices do not exist in the autarchic stage.

The autarchic state can come about in the economy because money has not yet been invented or the preconditions for its existence, such as trust, have not been established. As argued earlier in this chapter, the no-trade solution can also come about in a fiat money OLG model if the model has a finite horizon.

The autarchic solution (without money) of the fiat money OLG model can also come about if there is an expectation that the value of money will be zero in some future period. In equation (14), v_t depends upon the expected future value of money. That is, the value of money in the current period depends critically on the belief of the young that if they accept money in exchange for commodities in period t at the price p_t, they can exchange their money holdings for commodities, when old, in period $t+1$ at p^{e}_{t+1}. Suppose the young of period $t+i$, for any i, develop the contrary belief[22] that they cannot exchange their money holdings for commodities in period $t+i+1$ and that these holdings will become worthless. They will not be willing to accept money for commodities at any price in period $t+i$, so that the fiat money will have a zero value in period $t+i$, without any exchanges in period $t+i$ between the young and the old through the intermediation of fiat money. By backward extension, these results will also hold for all periods up to period $t+i$.

The autarchic solution (without money) is inferior to the stationary state with money since the latter allows each individual to rearrange his consumption over his lifetime on the basis of utility maximization, as against being limited in each lifestage to its endowment, which occurs in the autarchic case. Therefore, the contrivance of money increases welfare and has positive social value.

21.3.4 Indeterminacy of the price level and of the value of fiat money

Role of expectations and non-stationary states

Assume that the young in period $t+i$ expect a price p'_{t+i+1}, different from the price p_{t+i+1} determined by an appropriate version of equation (13). This will determine the price p'_{t+i} in period $t+i$, which in turn will determine that in $t+i-1$, and so on. There are, therefore, an infinite number of equilibrium prices between the expected future price p_{t+i} and zero, with the corresponding equilibrium values of money at which fiat money will be traded for commodities.

There are therefore two stationary solutions of the OLG model of fiat money. In one of these, the value of money is zero, money is not used and the individual does not trade with others. This is the no-trade or autarchic solution. The other stationary solution has a positive value of money; money is used and commodities are traded against money. The focus of the OLG models of money is on the latter. The value of money in this state is given by (13).

Fundamental solutions for prices, bubbles and tenuous equilibria

Given equation (13), designate the stationary paths of the commodity price and the value of money generated by the fundamentals of the system – i.e. by preferences, endowments and money supply – as being the "fundamental paths." The two stationary states belong to this category.

22 That is, contrary to the value implied by (13).

Since the current price depends on expected future prices, and the latter change often, the deviation thus produced of actual prices from the fundamental prices is said to be a "bubble" on the fundamental price path. This dependence of the current price on future prices in this argument – as incorporated in equations (13) and (5) – is known as *bootstrapping*: current prices are functions of the expected future prices; in the limit, if the latter were zero, the former would also be zero. This feature of the OLG models generates an infinite number of bootstrap paths – in which the path of the value of money depends on an arbitrarily specified expected value of money at a future date – with bubbles on the price level.

Given the bootstrap nature of the paths of the value of money in OLG models, none of these paths would be stable with respect to anticipated transitory shocks to the expected future value of money. It is claimed by some economists that this property reflects the "essential nature" of fiat money, that it is intrinsically useless, so that the amount an individual will pay for it in the current period depends critically upon the value that he expects to receive for it in the future. The indeterminacy and instability of monetary equilibria in OLG models have led to the monetary equilibria in these models being designated as *tenuous*.

In the non-stationary states, since the value of money need not necessarily be constant, the gross rate of return on it also need not be constant over time. However, any such state is unlikely to be stable and also would have a bootstrap path with bubbles on the price level. Studies of OLG models do not usually investigate non-stationary states, even though the common observation in the real-world economies is that of non-stationary states, in which prices, the value of money, the rate of return on money and expectations of their future values vary constantly. By focusing only on the stationary states, the standard OLG models rule out the study of these variations.

21.3.5 Competitive issue of money

Equation (13) implies that the equilibrium value of fiat money in the stationary state varies inversely with its quantity, so that this value will be positive only if the money supply is finite. Since such money – e.g. bank notes – is costless to produce, the competitive issue of fiat money would drive this value to zero. Hence, for a positive-valued equilibrium, OLG models rule out the competitive issue of fiat money by governments, as well as the issue of privately issued perfect substitutes that are virtually costless to produce.

21.4 The basic OLG model with a growing population[23]

To accommodate a growing population in the above OLG model, assume that:

$$N_{t+1}/N_t = n \tag{19}$$

where N_t is the number of persons born in t and n is the gross rate of increase in births. Therefore, the gross growth rate of population in this economy is given by:

$$[N_{t+2} + N_{t+1}]/[N_{t+1} + N_t] = [(n^2 + n)N_t]/[(n+1)N_t] = n$$

23 The following analysis can be easily adapted to that of a constant population but with endowments per person growing at a constant rate.

so that the gross growth rate of the population is also n. Now assume that each young individual in this economy continues to receive at birth an *ex gratia* amount w^y of commodities,[24] so that the total amount received increases proportionately with the population. From (13), the price level in this economy for the given money supply M_t is:

$$p_t = M_t/[N_t(w^y{}_t - c^y{}_t)] \tag{20}$$

so that, with the constant money supply ($M_t = M_{t+1} = m^o{}_t N_t$) in the economy and constant endowments per person at ($w^y - c^y$), the price level will evolve as:

$$p_{t+1}/p_t = N_t/N_{t+1} = 1/n \tag{21}$$

Hence the price of commodities will fall over time.

Rate of return on nominal balances when population is increasing

Given (21), the gross rate of increase in the value of money will be:

$$v_{t+1}/v_t = N_{t+1}/N_t = n \tag{22}$$

The gross rate of return r on nominal balances is the rate of increase in the value of money. It is given by:

$$r = n$$

Hence, nominal balances will have a positive gross rate of return equal to n and a net rate $(n - 1)$, determined by the growth rate of the population. In other words, in the current context of the endowment economy without production and a constant level of endowments per capita, the rate of return or "interest" on the only asset (i.e. money) in the model is determined by the biological growth rate.[25] If the model were to be extended to the case where the per capita endowments are also increasing at the rate η, the biological growth rate would have to be interpreted as equal to $(n \cdot \eta)$, with the gross return on money being equal to $(n \cdot \eta)$.[26]

24 A corresponding assumption is that each worker supplies just one unit of labor, with a constant marginal product per worker, and there exists full employment, so that the output of commodities increases at the same rate as the population.

25 This rate of return is also called the rate of interest, even though there are no loans in this model and this interest does not represent a distinct payment. Note that since the OLG models allow multiple solutions for the price level path, they also imply multiple solutions for the rate of interest. In one of these, the barter state, the rate of interest would be zero. However, the biologically determined rate of interest will be optimal. The social contrivance of money allows this rate to be implemented through market exchanges (Samuelson, 1958).

Note that in this simple model, there is no capital, so that the marginal productivity of capital cannot be defined.

26 If the economy's total endowments were constant but its population grew at the rate n, its per capita endowments would fall at the net rate $(n - 1)$.

21.5 Welfare in the basic OLG model

Welfare of the initial old

There are two distinctive sets of individuals (the initial old and the young) in the OLG model, so that we have to examine the welfare of each one. The initial old were allocated the newly created fiat money and did not at any time have to part with any commodities to get money balances, but received them *ex gratia*. Trade with the contrivance of fiat money enables them to exchange the money for commodities from the young. Therefore, there is a net benefit from trade with fiat money to the initial old.

Welfare of the initial young

In a comparison of the two stationary (trade and autarchic) solutions, each young individual – including those born in period t and after t – benefits from the trade with fiat money: it increases the overall utility that he can attain by being able to arrange a preferred consumption rather than having to consume the commodities according to the no-trade autarchic pattern. We prove this for *stationary equilibrium* with a growing population, a given money stock and the price ratio given by the market and derived in equation (21).

The economy's constraint for each period is that the aggregate amount of the commodities consumed in the period cannot exceed the aggregate stock of commodities. That is, in period t,

$$N_t c^{y}{}_t + N_{t-1} c^{o}{}_t \leq N_t w^{y}{}_t + N_{t-1} w^{o}{}_t \tag{23}$$

Dividing through by N_t yields:

$$c^{y}{}_t + (1/n) c^{o}{}_t \leq w^{y}{}_t + (1/n) w^{o}{}_t \tag{24}$$

Dropping the subscript t for the stationary state and rewriting (24) as an equality to focus on the economy's consumption frontier, we have:

$$c^{y} + (1/n) c^{o} = w^{y} + (1/n) w^{o} \tag{25}$$

Equation (25) is the per capita version of the economy's constraint and specifies the "economy's exchange frontier" or "socially feasible consumption tradeoff" between c^{y} and c^{o} in the stationary state. This is drawn as the line $W^{y}W^{o}$ in Figure 21.1. That is, with c^{y} on the horizontal axis, when c^{o} is zero c^{y} will equal W^{y}, where

$$W^{y} = w^{y} + (1/n) w^{o} \tag{26}$$

Similarly, when $c^{y} = 0$, c^{o} will equal W^{o}, where:

$$W^{o} = n(w^{y} + (1/n) w^{o}) \tag{27}$$

From (25), the socially feasible tradeoff between consumption when young and when old is given by:

$$\partial c^{o} / \partial c^{y} = -n \tag{28}$$

Now, turning to the personal lifetime budget constraint of each individual in the young lifestage, this is:

$$c^{y}_{t} + (p_{t+1}/p_t)c^{o}_{t+1} = w^{y}_{t} + (p_{t+1}/p_t)w^{o}_{t+1} \tag{29}$$

From (21), in the stationary equilibrium state with trade, the economy generates the price ratio $1/n$ for the case of stable money supply and population growth rate n. Hence we have, from (21) and (29),

$$c^{y}_{t} + (1/n)c^{o}_{t+1} = w^{y}_{t} + (1/n)w^{o}_{t+1} \tag{30}$$

In the stationary state with trade, with $c^{o}_{t+1} = c^{o}_{t}$ and $w^{o}_{t+1} = w^{o}_{t}$, and dropping the time subscript, (30) becomes:

$$c^{y} + (1/n)c^{o} = w^{y} + (1/n)w^{o} \tag{31}$$

Equations (31) and (25) are identical, so that the young individual's budget constraint is identical with the economy's per capita constraint. Hence, for both, the tradeoff between the consumptions of the young and the old is given by the line $W^{y}W^{o}$ in Figure 21.1. The slope $(\partial c^{o}/\partial c^{y})$ of the personal budget line $W^{y}W^{o}$ is $-n$, implying that since prices fall at the gross rate n, n units of the commodity can be purchased when old for one unit exchanged for money while young. Figure 21.1 also shows the individual's indifference curves I, I' and I'' as those specified by the individual's utility function (4). The young will maximize lifetime utility by being on the indifference curve that is tangential to the budget line $W^{y}W^{o}$. The optimal amounts consumed are given by the point B and are c^{y*} when young and c^{o*} when old.

Since the budget line and the economy's per capita constraint are identical, consumption at point B is feasible and maximizes the social welfare of the young generation over its lifetime. The level of utility at point B is higher than the level of utility obtained by the young from the autarchy (no trade) consumption of the original endowments w^{y} and w^{o}[27]

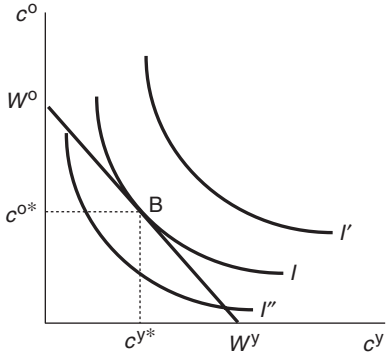


Figure 21.1

27 If $p_{t+1} = 0$, the young will not find it worthwhile to hold and carry money to the next lifestage, so that we would have a socially feasible frontier consisting only of the points w^{y} and w^{o}.

and also higher than from the trade pattern associated with any other point on the economy's exchange frontier $W^{y}W^{o}$. Therefore, trade at the equilibrium value of money given by (22) maximizes the intertemporal utility function of the young of generation t – as well as those of each succeeding generation.[28] Hence, trade and consumption given by point B is Pareto optimal, with Pareto optimality being defined as the maximization of the intertemporal utility of all generations (except the initial old, who have a special utility function of their own).

21.6 The basic OLG model with money supply *growth* and a *growing* population

Assume that the money supply increases at the gross rate θ – with the net rate equal to $(\theta - 1)$ – and population grows at the gross rate n. Therefore, the gross rate of increase in the money supply is specified by:

$$M_t/M_{t-1} = \theta \tag{32}$$

The new money supply in each period is assumed to be introduced into the economy by a free lump-sum gift of an identical amount to each of the old. Therefore, in each period, each old individual will have an amount that he had acquired in the preceding period through an exchange for commodities plus the lump-sum gift of newly created money. The change in money balances per old person is given by:

$$\begin{aligned}[M_t - M_{t-1}]/N^{o}{}_t &= (\theta - 1)M_{t-1}/N_{t-1} \\ &= (\theta - 1)m_{t-1}\end{aligned} \tag{33}$$

where the money balances per (young) person born in period $t-1$ (i.e. per old person in t) equal m_{t-1} and $(\theta - 1)$ is the net rate of growth of the money supply.

Determination of the commodity price

From (13), the equilibrium price levels in periods t and $t+1$ respectively are:

$$p_t = M_t/[N_t(w^{y}{}_t - c^{y}{}_t)] \tag{34}$$

$$p_{t+1} = M_{t+1}/[N_{t+1}(w^{y}{}_{t+1} - c^{y}{}_{t+1})] \tag{35}$$

In a stationary *real* equilibrium, $w^{y}{}_{t+1} = w^{y}{}_t$ and $c^{y}{}_{t+1} = c^{y}{}_t$, so that (35) can be rewritten as:

$$p_{t+1} = M_{t+1}/[N_{t+1}(w^{y}{}_t - c^{y}{}_t)] \tag{36}$$

28 Note that it does not maximize the utility of the old – in their old period – of any generation. Maximizing this utility means maximizing consumption in the old period, with their consumption while young being in the past and therefore without any tradeoff in consumption between the young and old lifestages.

Hence, from (32), (34) and (36), the gross rate of change of the price of the commodity is:

$$p_{t+1}/p_t = [M_{t+1}/M_t]/[N_{t+1}/N_t]$$

$$= \theta/n \qquad (37)^{29}$$

Rate of return on nominal balances in a growing economy with monetary expansion

From (37), the gross rate of change in the value of money is:

$$r_t = v_{t+1}/v_t = n/\theta \qquad (38)$$

Hence, for positive values of n and θ, (38) yields the following three cases.

1 If $\theta = n$, so that the money supply grows at the same gross rate as the population and the aggregate supply of commodities, the real value per unit of money will remain constant. In this case, the real value of the aggregate money supply will increase over time at the rate n.
2 If $\theta > n$, the rate of inflation will be $(\theta/n - 1)$ and the real value of the money supply will fall at the gross rate n/θ. In the special case when $n = 1$ (zero growth in population) but $\theta > 1$ (monetary expansion), the population and the supply of the commodity, as well as saving, will be constant but the money stock will grow at the rate θ. In this case the price level will increase at the gross rate θ, which is the same as the gross rate of increase θ in the money stock, implying that while the real value (per unit) of money will decline at the gross rate $1/\theta$, the real value of the aggregate money supply will remain unchanged over time.
3 If $\theta < n$, the rate of monetary expansion will be less than the growth rate of the population and commodities, so that the economy will experience a deflation of prices at the rate $(1 - \theta/n)$. The real value of the money supply will increase over time at the gross rate n/θ.

These conclusions are consistent with the Quantity Theory.

21.7 Inefficiency of monetary expansion in the money transfer case

We have already examined in Figure 21.1 the efficiency of a constant money supply, with population growing at the gross rate n. The present section shows the inefficiency in OLG models of money growth (or decline). Under the current assumptions, such money growth is achieved through a gratuitous lump-sum gift to members of the old generation, so that this case, as compared with others to be analyzed later in Chapter 22, will be labeled the *money transfer case*. The analysis in this section is again that of the stationary state.

29 Note that in (37), n is not only the rate of growth of population, it is also the rate of growth of output and saving in real terms. Hence, the rates of growth in real saving of n and in nominal money – in which saving is held – of θ will increase prices at the rate θ/n.

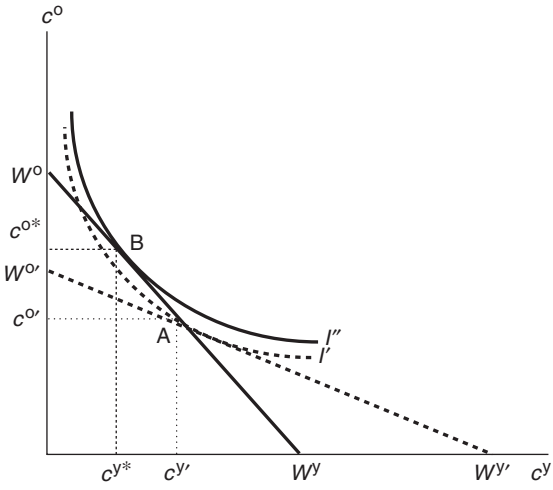


Figure 21.2

The economy's availability of commodities is not affected by the creation of money, so that the economy's feasible tradeoff (on a per capita basis) between c^o and c^y for the stationary state is not affected by the growth of the money supply and is as specified by equation (25). This was:

$$c^y + (1/n)c^o = w^y + (1/n)w^o$$

This tradeoff is shown by the line $W^y W^o$ in Figure 21.2, which is identical with the corresponding line in Figure 21.1. Note that the growth rate of the money supply does not enter (25), so that changes in it do not shift the economy's consumption frontier.

Now assume that money supply growth occurs over time at the gross rate θ, $\theta > 1$. The first and second periods' budget constraints of the individual become:

$$p_t c^y{}_t + m_t = p_t w^y{}_t \tag{39}$$

$$p_{t+1} c^o{}_{t+1} = p_{t+1} w^o{}_{t+1} + m_t + [M_{t+1} - M_t]/N^o{}_{t+1} \tag{40}$$

where the second term m_t on the right-hand side is the amount of saving carried through in money balances by the individual and the last term on the right-hand side is the receipt of a free lump-sum amount of newly created money in $(t+1)$. Substituting (39) in (40) to eliminate m_t gives the individual's lifetime budget constraint as:

$$c^y{}_t + (p_{t+1}/p_t)c^o{}_{t+1} = w^y{}_t + (p_{t+1}/p_t)w^o{}_{t+1} + (1/p_t)[M_{t+1} - M_t]/N^o{}_{t+1} \tag{41}$$

In this economy, $M_{t+1}/M_t = \theta, N^o{}_{t+1}/N_t = n$ and $M_t/N_t = m_t$. From (37), $p_{t+1}/p_t = \theta/n$, so that (41) can be restated as:

$$\begin{aligned} c^y{}_t + (\theta/n)c^o{}_{t+1} &= w^y{}_t + (\theta/n)w^o{}_{t+1} + (1/p_t)[(\theta-1)(M_t/N_t)] \\ &= w^y{}_t + (\theta/n)w^o{}_{t+1} + (1/p_t)[(\theta-1)m_t] \end{aligned} \tag{42}$$

In the stationary state, with the subscripts eliminated, the young individual's lifetime budget constraint (42) becomes:

$$c^{y} + (\theta/n)c^{o} = w^{y} + (\theta/n)w^{o} + (1/p)[(\theta - 1)m] \tag{43}$$

From the right-hand side of (43), the individual's intertemporal budget line in Figure 21.2 is $W^{y\prime}W^{o\prime}$, where:

$$W^{y\prime} = w^{y} + (\theta/n)w^{o} + (1/p)[(\theta - 1)m] \tag{44}$$

and

$$W^{o\prime} = (n/\theta)[w^{y} + (\theta/n)w^{o} + (1/p)(\theta - 1)m] \tag{45}$$

$W^{y\prime}$ and $W^{o\prime}$ are the maximum amounts that can be consumed respectively in the young and the old lifestages. Compared with the case when $\theta = 1$, the maximum possible consumption for $\theta > 1$ of the young increases because of the expected gratuitous receipt of money in old age, whereas the maximum possible consumption in the old lifestage falls, since the saving held in the form of money loses purchasing power through inflation caused by this receipt of the newly created money.

From (43), the slope of the young individual's lifetime budget line $W^{y\prime}W^{o\prime}$ in Figure 21.2 is:

$$\partial c^{o}/\partial c^{y} = -n/\theta \tag{46}$$

The economy's per capita consumption frontier (see equation (25)) defines the feasible combinations of c^{o} and c^{y} as $W^{y}W^{o}$ in Figure 21.2, with a slope of:

$$\partial c^{o}/\partial c^{y} = -n \tag{47}$$

Hence, for $\theta > 1$, the socially feasible tradeoff is steeper than the personal lifetime budget line in Figure 21.2.

Let the intersection point between the personal lifetime budget line and the socially feasible frontier in Figure 21.2 be the point A. The young's optimization requires him to choose a point on the budget line at which it is tangential to his highest indifference curve. The young cannot on average choose a point on the segment $AW^{y\prime}$ (excluding the point A itself) since all points on it are outside the feasible set and are therefore unattainable. The utility maximizing point can only be on the segment $AW^{o\prime}$ (inclusive of the point A), which is feasible from the economy's standpoint.

Now consider the points on the segment $AW^{o\prime}$. All points on $AW^{o\prime}$ other than A itself are inferior to points on AW^{o}: more consumption can be obtained for both young and old along AW^{o}, so that choosing a point along $AW^{o\prime}$ involves a welfare loss.

If the utility maximizing point was at A itself, as shown by the tangency of the budget line to the indifference curve I' through A, the indifference curve would have a slope of $-n/\theta$. This slope is less than that of the socially feasible tradeoff $W^{y}W^{o}$, so that I' must cut AW^{o} from below and there would be a higher indifference curve (say I'') which is tangential to $W^{y}W^{o}$ (at B) and gives greater intertemporal utility to the young, whose indifference curves were drawn as I' and I''. B would be a point on the segment AW^{o} and yields a higher level of utility to the young. Comparing points A ($\theta > 1$) and B ($\theta = 1$), "too much" is consumed while young and "too little" is saved for old age at point A relative to point B.

Further, the initial old would also benefit from moving from any point on $AW^{o\prime}$ to AW^{o}, since they would receive a higher level of consumption.

Therefore, both the young and the old suffer a welfare loss because of the divergence between the personal budget line and the socially feasible tradeoff. We have shown this for $\theta > 1$ but it also occurs for $\theta < 1$. For the money deflation case with $\theta < 1$ (not shown diagrammatically), the budget line would be steeper than the socially feasible tradeoff and, by reasoning similar to that for $\theta > 1$, there will be a welfare loss compared with the stable money case. In this case, too little will be consumed while young and too much saved for old age, relative to $\theta = 1$.

Welfare losses from changes in the money supply

To summarize, if the money supply in the economy were constant, θ would be equal to 1. In this case, the personal budget line and the socially feasible tradeoff would become identical and the individual's personal budget line would be the socially feasible tradeoff $W^{y}W^{o}$. The individual would maximize his utility at a point, say B, which is other than A and which lies on the segment AW^{o}, with B offering greater utility than A. Since the segment AW^{o} offers socially feasible consumption opportunities not available for $\theta > 1$, the individual would have a higher level of utility for $\theta = 1$ than for $\theta > 1$. The individual is therefore worse off with monetary expansion than with monetary stability.[30] Hence, monetary expansions and contractions imply a welfare loss for each individual and, therefore, for the economy as a whole.

Comparing intuitively the consumption bundles at points A and B in Figure 21.2 for $\theta > 1$, under monetary expansion, too much is consumed while young and too little is consumed while old. Consumption in the old lifestage is lower because its relative cost is higher than under monetary stability; given the population growth rate n, this relative cost under monetary expansion is θ/n whereas it is $1/n$ under monetary stability.

From another perspective, monetary expansion brings about a divergence between the social and private rates of exchange between c^{y} and c^{o}. The social rate of exchange $(\partial c^{y}/\partial c^{o})$ between c^{y} and c^{o} from the socially feasible tradeoff is $1/n$. However, under monetary expansion the private rate of exchange at which the individual is able to trade in the market is θ/n. The latter is higher since $\theta > 1$, so that the monetary expansion increases the private rate of exchange above the social rate and makes the individual choose too little of old-age consumption and too much of consumption when young. It does this by depreciating the value of money holdings, which lose their value at the rate θ compared with the case of monetary stability.

Note that on a per capita basis, the lifetime availability of commodities and therefore lifetime consumption is held constant in this model. Thus the welfare loss of monetary expansion is due not to a decrease in lifetime consumption but to an *inefficient* pattern of consumption. In the case of monetary expansion, too little is being consumed in old age because of the implied loss over time in the purchasing power of money, which is held for increasing old-age consumption but is not required for consumption while young.

30 The initial old who receive the increase in the money supply are also worse off, since they would be able to buy only $c^{o\prime}$ (at point A) of the commodity rather than c^{o*} (at point B) without monetary expansion.

Welfare loss from changes in the velocity of circulation

An alternative to monetary expansion is increases in the velocity of circulation of money. Velocity is constant at unity in this model, so that innovations in the payments mechanism that would change this velocity are not allowed in it. But if money in the economy were public (government supplied) fiat money M plus private fiat money M', with the latter, say, determined by $M' = \alpha M$, an increase in α would increase the velocity of fiat money. Since this corresponds to an increase in the monetary aggregate $(M + M')$, the increased velocity of public fiat money would imply a welfare loss.

21.8 Inefficiency of price stability with monetary expansion and population growth

Given a population growth rate of n, price stability could be achieved by a money supply rule that sets θ equal to n. With such a rule, prices would be stable but the monetary expansion would still impose an allocative inefficiency and a welfare loss on the economy. To see this for the stationary state, again start with the per capita socially feasible consumption frontier given by (25) for the money transfer case. This was:

$$c^{y}{}_{t} + (1/n)c^{o}{}_{t} = w^{y}{}_{t} + (1/n)w^{o}{}_{t}$$

The lifetime budget constraint from (43), with price stability ensured by $\theta = n$, would be:

$$c^{y}{}_{t} + c^{o}{}_{t} = w^{y}{}_{t} + w^{o}{}_{t+1} + (1/p)[(\theta - 1)m] \tag{48}$$

so that, for $\theta > 1$, the personal budget line is again flatter than the socially feasible tradeoff. As argued earlier in a similar analysis using Figure 21.2, the consumption pattern chosen by the young would be inefficient and would involve an allocative welfare loss for both the young and the old.

Therefore, the allocative inefficiency is that of monetary expansion (or contraction) rather than of inflation (or deflation). This inefficiency again flows from the divergence created by monetary expansion between the social rate of exchange between c^{o} and c^{y} and the private one. Population growth makes this rate n – so that n units of c^{y} can be traded for c^{o} – whereas the present money supply rule with $\theta = n$ makes it 1.

We leave it to the reader to show that a negative population growth rate with $n < 1$ (the declining population case) but a stable money supply will cause inflation but not impose a welfare loss.

21.9 Money demand in the OLG model with a positive rate of time preference

We have so far not looked at the specification of the intertemporal utility function, though we derived in (46) the result that the individual's equilibrium intertemporal rate of substitution between c^{o} and c^{y} would be $-n/\theta$. This section applies the above analysis to a specific time-separable utility function with a positive rate of time preference. Doing so allows the derivation of the c^{y}, c^{o}, s and m^{d} functions.

A common assumption on the form of the intertemporal utility function is that it is time separable in period utility functions and that period utility is the log of consumption in the period. That is:

$$U(c^{y}_{t}, c^{o}_{t+1}) = u(c^{y}_{t}) + \delta u(c^{o}_{t+1}) \tag{49}$$

$$= \ln c^{y}_{t} + \delta \ln c^{o}_{t+1} \tag{50}$$

where:

$U(.)$ = intertemporal utility function
$u(.)$ = period utility function
δ = gross subjective discount factor (1 divided by the gross rate of time preference).

Note that the assumed utility function does not directly or indirectly include money holdings as an argument, thereby meeting the requirement of the benchmark OLG models of fiat money that money is "intrinsically useless."

The appropriate lifetime budget constraint for the case of population growth at the rate n and monetary growth at the rate θ is given by (42) as:

$$c^{y}_{t} + (\theta/n)c^{o}_{t+1} = w^{y}_{t} + (\theta/n)w^{o}_{t+1} + (1/p_t)[(\theta - 1)m_t]$$

The right-hand side of this constraint can be designated as current wealth W_t, so that:

$$W_t = [w^{y}_{t} + (\theta/n)w^{o}_{t+1} + (1/p_t)(\theta - 1)m_t]$$

Maximizing (50) subject to (42) yields the optimal consumption pattern as:

$$c^{y}_{t} = [1/(1+\delta)]W_t \tag{51}$$

$$c^{o}_{t+1} = (\delta n/\theta)c^{y}_{t} = [\delta n/(\theta(1+\delta)]W_t \tag{52}$$

The individual's saving in period t is:

$$s_t = w^{y}_{t} - c^{y}_{t}$$

$$= w^{y}_{t} - [1/(1+\delta)]W_t \tag{53}$$

Demand for money in the OLG model

The individual's demand for real balances m^{d}_{t}/p_t is positive only in the young lifestage and occurs as a way of carrying forward his saving. It equals his saving and is given by:

$$m^{dy}_{t}/p_t = w^{y}_{t} - [1/(1+\delta)]W_t$$

or as:

$$m^{dy}_{t}/p_t = w^{y}_{t} - [1/(1+\delta)][w^{y}_{t} + (\theta/n)w^{o}_{t+1} + (1/p_t)(\theta - 1)m_t] \tag{54}$$

Note that, in the old lifestage, the individual has negative saving, so that his money demand is zero.[31] Also note that since $m^{d}{}_{t} = m^{dy}{}_{t} + m^{do}{}_{t}$ and $m^{do}{}_{t} = 0$, $m^{d}{}_{t} = m^{dy}{}_{t}$, so that all of the money demand comes from the young. Further, $m^{dy}{}_{t} = s^{y}{}_{t}$. Hence, money demand is determined only by the saving of the young.

In the case of *price stability*, i.e. with $\theta = n$, we get:

$$c^{o}{}_{t+1} = \delta c^{y}{}_{t} \tag{55}$$

$$m^{d}{}_{t}/p_{t} = w^{y}{}_{t} - c^{y}{}_{t} = w^{y}{}_{t} - [1/(1+\delta)]W_{t} \tag{56}$$

Hence, the demand for real balances – and saving – depends positively upon current income $w^{y}{}_{t}$ and the gross subjective discount factor δ, but negatively upon future income $w^{o}{}_{t+1}$.

We leave it to the reader to derive the price level for the above exercise and show the superiority of using money to barter, as well as the inefficiency of money growth.

21.10 Several fiat monies

In the preceding version of the OLG model with a single type of fiat money, the individual used fiat money to transfer purchasing power between his lifestages. As shown above, fiat money had a gross rate of return n/θ for a population growth rate n and a monetary expansion rate θ. This return was in the form of a "dividend" or "tax" occurring through changes in the price level over time, even though the fiat money itself did not make an explicit interest payment.

If there are several fiat monies simultaneously available in the economy, and each one only offers the services of a store of value, the one with the highest return (coupon plus the rate of increase in real value) will be preferred by the savers and the fiat monies with lower rates of return will not be held. Conversely, if several fiat monies offer the same rate of return, the savers will be indifferent among them and the relevant rate of monetary expansion will be that of all such fiat monies taken together.

Applying these results to the international context of many national currencies and open economies with zero transactions/exchange costs between currencies, either only one currency would possess a positive demand and positive value – in the absence of legal restrictions or frictions favoring the domestic currency – or all national currencies would be perfect substitutes for each other. In the latter case, with seigniorage from the national currencies going to their home nations, *profit/seigniorage maximizing* governments would engage in a competitive issue of fiat monies until the value of all fiat monies went to zero. Therefore, perfectly substitutable and competitively issued national fiat monies do not make sense in the context of frictionless OLG models. Alternatively, if many such monies exist in the real world, then the OLG models would not be valid for them or the positive demand for several currencies, including the domestic one, simultaneously will only arise because of legal and other restrictions on the use of foreign currencies.

In the real world, national currencies usually remain the most commonly used fiat money in their own nation, even though they tend to have different rates of inflation and return

31 This is, of course, clearly not valid in the monetary economy in which everyone (including those in their last year of life) who wishes to obtain commodities must purchase them against money, so that everyone would hold some money on average over each period of their life. This is so whether saving during the period is positive, zero or negative.

from foreign currencies. The OLG models imply that this national usage must be due to legal restrictions or other frictions that prevent perfect substitutability between the national fiat money and foreign ones, thereby creating special niches or market segments in which each can dominate. The domestic demand for the national currency is to be determined by the nature and extent of frictions and restrictions, which are themselves often not specified or encompassed in the OLG models. To take a concrete example, consider the Canadian economy in which there do not exist any legal barriers to the use of US currency relative to Canadian one. Both circulate freely, payments at many stores can be made in either currency, deposits in banks can be held in either currency, etc., and, between 1980 and 2003, the expectation that the US dollar will appreciate relative to the Canadian one was often held. Yet, the Canadian currency, rather than disappearing from circulation, continued to be the dominant one in use in Canada. Nor was there any significant difference in relative use between the Canadian and American currencies in the 1980–2003 period versus the period 2003–2005, when the Canadian dollar appreciated strongly relative to the US dollar.

21.11 Sunspots, bubbles and market fundamentals in OLG analysis

The OLG models with positively valued fiat money are said to display "sunspot" activity. Sunspots are variables which are not among the fundamental variables of the economy but which can nevertheless affect the economy. The usual examples of fundamental variables in the OLG literature are endowments, preferences, technology and market structures. An example of a sunspot is the market expectation about the future values of a variable that influences its present value. As shown in this chapter, expectations do play such a role with respect to the value of fiat money in the OLG models, so that these models are very prone to sunspot activity. The common limitation of the analysis of such models to stationary equilibrium states is a means of avoiding sunspots. In such stationary states, the value of money is determined by the fundamental determinants of the economy, with rational expectations specifying the expected value of money to be that prescribed only by the stationary state values.

A related concept to that of sunspots is that of "bubbles" in asset prices. A price bubble for an asset exists if its market value is different from the present discounted value of its expected dividends. The latter is designated as its market fundamental. Since fiat money in the OLG models does not yield any implicit (such as in terms of utility or productivity) or explicit return (such as a coupon or dividend payment) to its holder, its fundamental value is zero. Since it can have a positive value in these models, its market price can differ from its market fundamental, so that there is a bubble in positively valued money. Such bubbles can be created by sunspots. Further, bubbles on an asset can have nominal and real consequences for other variables in the model. In the OLG models, the actual bubble on the value of fiat money has consequences for the consumption and saving path over time.

Sunspots and bubbles are defined relative to the outcomes based on market fundamentals, which are defined in the OLG literature as endowments, technology and preferences. However, the exclusion from this list of the financial structure of the economy is incorrect for a monetary economy. The financial system is one of the major structures of the economy, just as other sectors such as energy, transport, etc., are. It uses resources and produces output, and its efficiency is critical to the production levels of other industries and social welfare. For a monetary economy, it is not a superficial imposition on an otherwise barter economy, but needs to be listed and treated as one of the market fundamentals. Consequently, a positive

value of money is not really a sunspot or a bubble on its "fundamental" value, as asserted by the benchmark OLG model. In fact, a positive value of money is a stylized fact of all monetary economies, and not an oddity among such economies. The fundamental role of money and the financial system in the economy is further discussed in Chapter 24 on monetary growth theory.

Conclusions

The OLG framework offers an alternative format to MIUF and MIPF theories (in which economic agents have an infinite horizon) for modeling the demand for money. They are especially appealing to those who believe that, since money cannot be directly consumed or used as a factor of production, it should not be entered as a variable in the utility and production functions. The OLG models do not do so, but still generate a demand and positive value for fiat money, even though it is intrinsically costless to create. However, to ensure a positive value of money, they must impose a restriction on its supply, so that the OLG models rule out the competitive supply of fiat money.

In comparison with analyses using agents with an infinite horizon, the OLG framework is a vehicle for the intertemporal reallocation of resources to the future by individuals with finite lifetimes but operating in an everlasting economy. It provides a market mechanism as a substitute for an inter-generational social contract (to provide transfers of commodities to the elderly who are no longer productive but who, while young, made similar transfers). Limited elements of an implicit social contract continue to exist: the faith (or expectation) is firmly held that no future generation will make the fiat money inconvertible into commodities directly or indirectly by the substitution of a new currency which renders the preceding one inconvertible into the new one.

Further, the current value of fiat money will depend upon what the future generations are expected to maintain, so that the OLG models show multiple equilibria for the value of fiat money. Of these, only two relate to a stationary state. These are the no-trade (autarchic) equilibrium and an equilibrium in which money has a positive value.

Since the balances of fiat money are a durable or capital good, the value of such money in any period will depend upon its value in the next period, which will depend upon its value in the period after that, and so on. In the general case, an expected change in the future value of money in any future time period will change its value in all preceding periods. Therefore, the current value of money and the current price level in the economy are tenuous, and can involve a bubble on the value of money. While this is so for most intertemporal theories, including the MIUF theory, the OLG models additionally assert that if any of these future values are expected to be zero, all preceding values would also be zero. This is not surprising since the fundamental assumption of the OLG theory is that fiat money possesses a one-dimensional use, i.e. it is a store-of-value function for carrying purchasing power to the future – a use that is totally eliminated if its expected future value is zero.

A major concern of this book, as of monetary economics generally, is the theoretical derivation of a money demand function that is empirically valid. As shown in this chapter, the demand for real balances in the basic OLG model is specified by saving. Such a function is empirically testable and is, without dispute, not valid for any real-world economy. This departure from the empirical money demand functions is so radical that it casts serious doubt on the basic OLG model as a satisfactory basis for monetary economics. This issue is discussed further in the next chapter.

The relationship between the demand for money and saving in the OLG model highlights its basic deficiency: it does not provide a medium-of-payments role for money, as Tobin (1980) and McCallum (1983) emphasize.

Summary of critical conclusions

- ❖ OLG models incorporate a close relationship between the demand for money and saving, rather than between the demand for money and current consumption expenditures.
- ❖ OLG models imply multiple equilibria for the price level and the return on money but only two stationary equilibria, one of which is the no-trade autarchic state in which money is not traded, while the other is one in which money is used and has positive value.
- ❖ A positive value of fiat money represents a bubble since fiat money is costless to create.
- ❖ According to the basic OLG model, monetary expansion or contraction yields a less desirable allocation of consumption in the young and old lifestages relative to a constant money supply.
- ❖ According to the basic OLG model, while inflation caused by monetary expansion – and disinflation caused by monetary contraction – involves a welfare loss because it is the outcome of monetary expansion, inflation caused by a fall in commodity endowments or population decrease, with a constant money supply, does not do so.
- ❖ According to the benchmark OLG model, in the presence of population growth, price stability achieved through monetary expansion is inefficient relative to a price decrease under a constant money supply.
- ❖ The basic OLG model does poorly in generating implications consistent with the stylized empirical facts about money in the economy.

Review and discussion questions

1. Specify and discuss at least five stylized facts about money in the economy.
2. For the benchmark OLG model, show and discuss the welfare implications of monetary expansion that ensures price stability, along with those of monetary stability that results in price deflation.
3. For the OLG model, you are given the utility function of the representative young as:

$$U(.) = \ln c^{y}_{t} + \delta \ln c^{o}_{t+1}$$

   For a given population and given endowments of the commodity in the two periods, derive the demand functions for the commodity, the price level and the rate of return on nominal balances in the stationary state.
4. In the OLG framework, assume that the economy has N persons born each period, each person lives two periods, each young person supplies one unit of labor and saves a constant proportion of income. The old do not supply labor, nor do they save. During each period the economy's saving can be held in physical capital, which can be bought at the end of the young lifestage and lasts only one period (i.e. during the old lifestage), or fiat money. The economy has the production function:

$$y = f(k) = Ak^{\alpha} \quad 0 < \alpha < 1$$

where k is the capital/labor ratio. Derive the economy's steady-state output, saving, the capital stock and the demand for money functions. (Specify any additional assumptions that you need to make.) Is money neutral in this model?

5. Experience indicates that in the real world, several fiat monies – for example, the Canadian and US dollars in Canada – can coexist in the economy with positive values even when there are no legal barriers to the use of each one. Further, each has its own determinate demand function, quite different from the other's. How do you reconcile this empirical fact with the analysis of the OLG model in this chapter, or are they irreconcilable without fundamental alterations in the model?
6. For the fiat money in your country, what would be the arguments and signs of the partial derivatives of its demand function, according to: (a) the OLG model of this chapter; (b) your knowledge of the economy and the empirical literature on it?
7. In the OLG framework, assume that the incomes in the (one-commodity) economy are paid to workers in the form of the commodity by the representative private firm, with full employment in the economy. Assume that each worker produces one unit of the commodity (without requiring physical capital). Specifying any other assumptions that you need to make, derive the implications for prices, the return on money, the efficiency of monetary expansion and the optimal rate of monetary expansion for the following cases:

   (i) a constant labor force;
   (ii) the labor force grows at the rate n;
   (iii) the labor force is constant, but its average productivity rises at a constant rate τ over time.

8. The assumption in many, if not most, macroeconomic models is that firms do not save but borrow (through the issue of one-period bonds) each period the funds that they need for investment. Under this assumption, if fiat money is intrinsically useless for firms, its demand by firms will be zero under the rationale of the benchmark OLG model. How, then, can the use of money by firms be generated and be positive in the OLG context?
9. Discuss the statement: money will have no value if there is a terminal date for the economy.
10. Discuss the statement: the uniqueness of monetary equilibrium with perfect foresight in the basic OLG model with money requires implausible conditions.
11. Discuss in the context of the OLG model the validity of Friedman's money supply rule: the best monetary policy is one that *increases* the money supply at a steady low rate.
12. Given the stylized empirical facts about money in the economy, discuss the validity of the benchmark OLG model. If its implications are drastically at odds with the stylized empirical facts, should it be retained? If so, on what grounds?

References

Champ, B., and Freeman, S. *Modeling Monetary Economies*. New York: John Wiley and Sons, 1994, Chs. 1–3, 5, 6.

McCallum, B.T. "The role of overlapping generations models in monetary economics." *Carnegie-Rochester Series on Public Policy*, 18, 1983, pp. 9–44.

Samuelson, P.A. "An exact consumption-loan model of interest with or without the social contrivance of money." *Journal of Political Economy*, 66, 1958, pp. 467–82.

Sargent, T.J. *Dynamic Economic Theory*. Boston: Harvard University Press, 1987, Ch. 7.

Tobin, J. "Discussion." In J.H. Karekan, and N. Wallace, eds, *Models of Monetary Economics*. Federal Reserve Bank of Minneapolis, 1980, pp. 83–90.

Wallace, N. "The overlapping generations model of fiat money." In J.H. Karekan, and N. Wallace, eds, *Models of Monetary Economics*. Federal Reserve Bank of Minneapolis, 1980, pp. 49–82.

Wallace, N. "A Modigliani–Miller theorem for open market operations." *American Economic Review*, 71, 1981, pp. 267–74.

22 The OLG model

Seigniorage, bonds and the neutrality of fiat money

This chapter extends the analysis of the overlapping generations models of money to include bonds. It then examines various policy issues such as seigniorage, open market operations and money neutrality.

An important distinction between the different ways of increasing the money supply arises in the OLG models of this chapter. In the benchmark OLG model, open market operations have no impact on the economy, even to the extent of not changing the price level, which is clearly not valid for the modern economy. However, OLG models with labor supply or/and production often imply the non-neutrality of money.

This chapter also examines the empirical plausibility of important implications of the OLG models, especially for the demand for money, and rejects their validity.

Key concepts introduced in this chapter

- Seigniorage as a taxation device
- Open market operations in the OLG models
- The Wallace–Modigliani–Miller theorem on open market operations

This chapter continues the exposition of the various aspects of the basic OLG model with money. The assumptions of the model and the definitions of the symbols used are as in the preceding chapter. The present chapter has numerous references to the equations in Chapter 21, with such equations being cited as (21…). The stylized facts presented in Chapter 21 need to be kept in mind in evaluating the models presented in this chapter.

The chapter starts by considering seigniorage from monetary expansion in Section 22.1, with seigniorage as a taxation device. Section 22.2 introduces bonds into the OLG model and Section 22.3 presents Wallace's (1980) version of the Modigliani–Miller theorem for open market operations in the OLG model with money. This theorem, labeled the W–M–M theorem in this chapter and the next, implies that changes in the supply of fiat money through open market operations will not have any impact on other economic variables, including the price level and nominal national income, let alone the real variables of output and unemployment!

Section 22.4 presents an illustrative OLG model with money and production to explore the determination of the capital stock, output, prices and the value of money. Money proves not to be neutral in this model. Section 22.5 re-investigates the neutrality and non-neutrality of money in a slightly different model, and shows that neutrality or non-neutrality depends

on the structure of the economy and the assumptions on how money is introduced into the economy.

The test for the acceptance or rejection of any theory in economics is its ability to explain observed behavioral relationships. Section 22.6 addresses the fundamental question of whether or not the benchmark OLG model with fiat money explains the estimated forms of the money demand function and other major facets of the monetary economy. Since it does not do so, this chapter concludes with a call for a substantial modification of this model to include the medium-of-payments role of money, as well as to include interest-paying bonds, in a realistic manner. These will be undertaken in the next chapter.

22.1 Seigniorage from fiat money and its uses

This section assumes that the government uses newly created money to buy commodities from the private sector, so that money creation now generates seigniorage. Note that this assumption differs from that in the preceding chapter, which was that the newly created money is given gratuitously to the old, so that there was no net seigniorage accruing to the government from money creation.

Seigniorage is the revenue from monetary expansion.[1] The nominal value of the seigniorage per young person from monetary expansion at the gross rate θ in period t is given by:

$$\begin{aligned} G_t &= [M_t - M_{t-1}]/N_t \\ &= (1 - 1/\theta)M_t/N_t = (1 - \theta)m_t \end{aligned} \tag{1}$$

where G_t is the nominal value of seigniorage per person born in period t, $M_{t-1}/M_t = 1/\theta$ and $m_t = M_t/N_t$. $(1 - 1/\theta)$ is viewed as the "tax rate" and the money supply per capita (i.e. m_t) is viewed as the per capita "tax base." Define g_t as G_t/p_t, so that g_t is the amount of *real* seigniorage per young person in the economy. The amount of commodities that can be bought with this amount is also g_t, where g_t is:

$$g_t = (1 - 1/\theta)m_t/p_t \tag{2}$$

The seigniorage received by the government can be allocated by it in various ways. Among these, two distinctive forms are:

(i) The seigniorage is distributed by the government to the old members of each generation in the form of gratuitous money transfers and, therefore, as a unilateral transfer back from the public sector to the private one. This was the assumption used in Chapter 21.

1 To place the relevance of seigniorage in empirical perspective, estimates of seigniorage from money creation are often less than 2 percent of GDP but could be as high as 10 percent or more, depending on the extent of money creation and inflation; as a percentage of government expenditures, the estimates are less than 10 percent for most countries, but much higher for a few countries (see Click, 1998).

(ii) The seigniorage is used by the government to purchase commodities in the open market from the private sector, without a gratuitous return of the commodities back to the private sector.[2] This chapter makes this assumption.

In addition to (ii), this chapter assumes that in each period *the government uses its seigniorage to buy commodities and then destroys them* – or gives them away as a unilateral gift to foreigners (economic units outside the economy in question) – without producing any welfare for the economy through such action.[3] Such an assumption is of doubtful validity and is clearly irrational. Its purpose in such models is to isolate the effects of creating seigniorage from those arising from the distribution of the value of seigniorage, either in the form of money transfers or commodity transfers to the population or some segment of it, and to compare these effects with those of lump-sum taxation as one of the alternatives to seigniorage.[4]

22.1.1 Value of money under seigniorage with destruction of government-purchased commodities

Under the above assumptions for this case, the economy is left with $g_t N_t$ less of goods in period t. The economy's constraint, equation (21.23) in Chapter 21, for the commodity market in period t is now modified to:

$$N_t c^{y}{}_t + N_{t-1} c^{o}{}_t = N_t w^{y}{}_t + N_{t-1} w^{o}{}_t - N_t g_t$$

Let $N_t/N_{t-1} = n$, so that, as shown in Chapter 21, population grows at the gross rate n. Dividing the previous equation by N_{t-1}, we have:

$$c^{y}{}_t + (1/n) c^{o}{}_t = w^{y}{}_t + (1/n) w^{o}{}_t - g_t \tag{3}$$

The demand for real balances arises from the need of the young to carry purchasing power to the following period when they will be old. The saving of the young equals their endowment of commodities less their consumption. Therefore, the per capita demand for real balances, m^{d}/p_t, is given by:

$$m^{d}{}_t/p_t = w^{y}{}_t - c^{y}{}_t - g_t \tag{4}$$

The nominal supply of money is:

$$M^{s}{}_t = M_t \tag{5}$$

Equilibrium in the money market requires that:

$$p_t[N_t(w^{y}{}_t - c^{y}{}_t - g_t)] = M_t \tag{6}$$

2 Modern governments usually use the seigniorage transferred to them by the central bank to fund their expenditures, which are used to provide government services or goods to the public. This procedure makes (i) somewhat more realistic than (ii).

3 The usual usage of government-purchased goods and services is the provision of publicly provided goods to the economy. This is different from the assumption being made here.

4 Admittedly, the two are inextricably linked in this two-asset – commodity and money – economy, so that attempting to disentangle them through the assumption of the destruction of the commodity equal to seigniorage introduces a degree of irrationality and unrealism into the analysis.

so that:

$$p_t = M_t/[N_t(w^y{}_t - c^y{}_t - g_t)] \tag{7}$$

Equation (7) shows that an increase in seigniorage g raises the price level.

Since w^y, c^y and g will be constant in the stationary state but M_t and N_t are growing, (7) implies that, in the stationary state,

$$p_{t+1}/p_t = \theta/n \tag{8}$$

From (7), the value of money is:

$$v_t = [N_t(w^y{}_t - c^y{}_t - g_t)]/M_t \tag{9}$$

and, from (8), the rate of return on money is:

$$v_{t+1}/v_t = n/\theta \tag{10}$$

As (7), in comparison with equation (21.34) of Chapter 21, shows, the price level will be higher and the availability of commodities less in this case with the destruction of the government-acquired goods than with the gratuitous money transfer case. However, comparing (8) and (10) with equations (21.37) and (21.38) of Chapter 21 shows that the intertemporal price ratio and the real rate of return remain the same in the two cases.

22.1.2 Inefficiency of monetary expansion with seigniorage as a taxation device

Remember that in the case being analyzed in this section the government is assumed to destroy commodities equal to the seigniorage, rather than transferring them to the old within the same period as under the money transfer case in Chapter 21. Since the commodity destruction reduces the amount of commodities available to the economy, social welfare under this policy must be lower than under the money transfer policy. Further, since the latter was inferior to the stable money case, monetary expansion with commodity destruction must also mean a welfare loss over the stable money case. We prove this formally in the following.

We want to derive at this point the personal budget constraint for the monetary expansion case with commodity destruction. In this case, since there is no money transfer to the individual, monetary creation does not affect the individual's personal budget constraints for the two lifestages, nor does the destruction of the commodity acquired by the government through seigniorage. These personal per-period constraints remain identical to the ones without monetary expansion and are:

$$p_t c^y{}_t + m_t = p_t w^y{}_t \tag{11}$$

$$p_{t+1} c^o{}_{t+1} = p_{t+1} w^o{}_{t+1} + m_t \tag{12}$$[5]

5 Note the difference between (12) and equation (21.40) of Chapter 21, though both involve an identical monetary expansion.

Substituting (11) in (12) to eliminate m_t gives the personal lifetime budget constraint as:

$$c^{y}{}_{t} + (p_{t+1}/p_t)c^{o}{}_{t+1} = w^{y}{}_{t} + (p_{t+1}/p_t)w^{o}{}_{t+1} \tag{13}$$

Since $p_{t+1}/p_t = \theta/n$, (13) becomes:

$$c^{y}{}_{t} + (\theta/n)c^{o}{}_{t+1} = w^{y}{}_{t} + (\theta/n)w^{o}{}_{t+1} \tag{14}$$

Therefore, in stationary equilibrium,

$$c^{y} + (\theta/n)c^{o} = w^{y} + (\theta/n)w^{o} \tag{15}$$

Hence, the individual's personal budget line in Figure 22.1 is $W^{y\prime\prime}W^{o\prime\prime}$, where:

$$W^{y\prime\prime} = w^{y} + (\theta/n)w^{o} \tag{16}$$

$$W^{o\prime\prime} = (n/\theta)[w^{y} + (\theta/n)w^{o}] \tag{16$'$}$$

Since the seigniorage from monetary expansion is now no longer returned to the people, the personal budget line from (15) lies *inside* the one (not shown in Figure 22.1) from equation (21.43), so that welfare is even lower in the commodity destruction case than in the money transfer case, though both have the same slope $(-n/\theta)$. Hence, a policy of commodity destruction is inferior to one that transfers money or goods back to the public.

Further, in Figure 22.1, for the choices made by the young, comparing the young's personal budget line $W^{y\prime\prime}W^{o\prime\prime}$ and the economy's per capita constraint $W^{y}W^{o}$, the former has a flatter slope equal to $(-n/\theta)$, for $n, \theta > 1$, than the latter whose slope from (3) equals $(-n)$. Let the intersection point between these lines be at X. Similar to our arguments in Chapter 21 for the money transfer case, the segment $XW^{y\prime\prime}$ (excluding point X) is not feasible since it lies outside the economy's constraint, so that the representative young's utility-maximizing point will have to lie on the segment $XW^{o\prime\prime}$ of the personal budget line. But, as shown in

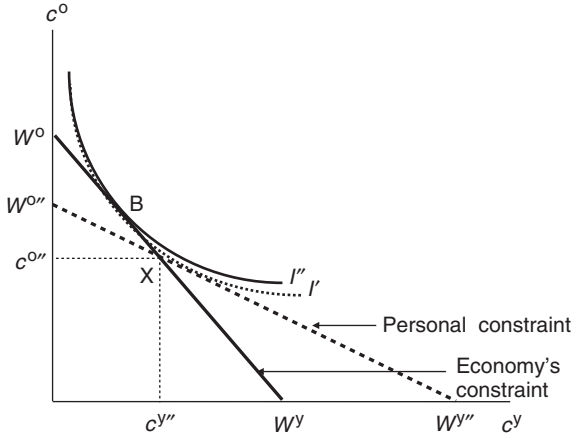


Figure 22.1

Figure 21.2 of Chapter 21, any point on this segment is inferior in utility to some points on the segment XW^{o} of the economy's constraint, with the latter offering more of both c^{y} and c^{o}. Since the segment XW^{o} is available to the young in the case of monetary stability (i.e. with $\theta = 1$), monetary expansion ($\theta > 1$) decreases the welfare of the young relative to that under monetary stability. Further, the consumption and welfare of the old would also be higher if the young were to choose a point on the segment XW^{o} of the economy's constraint. Hence, there is a social welfare loss from the seigniorage as a taxation *cum* commodity destruction device.

Monetary stability is therefore preferable to monetary expansion with money transfer to the old, which in turn is preferable to monetary expansion with destruction of goods bought with seigniorage.

22.1.3 Change in seigniorage with the rate of monetary expansion

From (2), the real value of the seigniorage per person collected from the economy is:

$$g_t = (1 - \theta^{-1})M_t/(p_t N_t) = (1 - \theta^{-1})m_t/p_t \tag{17}$$

where M_t/p_t represents the real tax base – that is, the amount subject to taxation – and $(1 - \theta^{-1})$ represents the tax rate on the real value of the money holdings. From (17),

$$\frac{\partial g_t}{\partial \theta} = \{\theta^{-2}\frac{m_t}{p_t}\} + (1 - \theta^{-1})\frac{\partial (m_t/p_t)}{\partial \theta} \tag{18}$$

As θ increases, the opportunity cost of holding real balances increases, so that $\partial(m_t/p_t)/\partial\theta < 0$. Hence, for $\theta > 1$, the first term on the right side of (18) is positive and the second term is negative. The first term causes an increase in seigniorage as θ increases while the second term causes a decrease as the tax base shrinks. If we assume that the second effect is initially weaker but gradually becomes stronger as θ increases, the amount of seigniorage as a function of θ would first increase and then decrease.[6]

22.1.4 Change in the lifetime consumption pattern with the rate of monetary expansion

From (11), ${c^{y}}_t = {w^{y}}_t - m_t/p_t$. Therefore, noting that $\partial(m_t/p_t)/\partial\theta < 0$,

$$\frac{\partial {c^{y}}_t}{\partial \theta} = -\frac{\partial (m_t/p_t)}{\partial \theta} > 0 \tag{19}$$

From (12), ${c^{o}}_t = {w^{o}}_t + (m_t/p_t)\,(p_t/p_{t+1})$, where $(p_t/p_{t+1}) = n/\theta$. Therefore,

$$\frac{\partial {c^{o}}_t}{\partial \theta} = \frac{n}{\theta}\frac{\partial (m_t/p_t)}{\partial \theta} - \frac{m_t}{p_t} n\theta^{-2} < 0 \tag{19'}$$

Intuitively, an increase in θ increases the tax rate on money balances in which saving by the young is held. Therefore, it reduces such saving, thereby increasing the consumption in the

6 $\partial^2 g/\partial\theta^2$ can be positive, zero or negative.

young lifestage and decreasing in the old lifestage. That is, increasing the rate of monetary expansion tilts consumption from the old to the young lifestage.

22.1.5 Seigniorage from monetary expansion versus lump-sum taxation

To compare the relative merits of monetary expansion as a taxation device with lump-sum taxation in the preceding framework, assume for the latter case that a lump-sum tax[7] equal to g_t is imposed on each *young* individual[8] in period t, *with the money supply now being held constant*. Note the underlying assumption on the destruction of the commodity bought with the seigniorage from money or with the tax g_t. The economy's constraint in this case is as specified in (3). The lifetime personal budget constraint in this case is a modified version of (14) with $\theta = 1$ and with g_t being deducted from the available resources on the right-hand side of (14). That is, the personal budget constraint becomes:

$$c^{y}{}_{t} + (1/n)c^{o}{}_{t+1} = w^{y}{}_{t} + (1/n)w^{o}{}_{t+1} - g_t \tag{20}$$

This personal budget constraint is identical with the economy's constraint (3), so that the consumption pattern (c^{y}, c^{o}) chosen by the individual also maximizes the welfare for the economy. Therefore, with the reduction in commodities by g being the same, lump-sum taxation is an efficient taxation device, while seigniorage from monetary expansion is not.

22.1.6 Seigniorage as a revenue collection device

The above comparison between seigniorage and lump-sum taxation for revenue generation assumes that the costs of collection and administration are identical between them and implies that lump-sum taxation is preferable. This is almost never so in the real world. The cost of monetary expansion through the printing of more money is negligible and the economy, through the money markets, takes care of the administration of seigniorage through intertemporal price changes without direct collection costs to the government. By comparison, the bureaucracy needed for the collection of lump-sum (as well as income) taxes, their visibility as a tax on the public and the consequent public resentment at their payment reduce society's welfare and reduce the amounts actually collected on both a gross and a net (of the cost of collection) basis. These costs can be high enough in certain economies to ensure that some monetary expansion may be preferable as the policy imposing the smallest possible welfare loss on the economy.[9] These considerations make the collection of revenue through seigniorage, in addition to direct and indirect taxes, tempting for governments in many countries. But, as against such advantages of seigniorage as a taxation device, monetary expansion and inflation have other costs – such as those related to the variability of inflation,

7 A lump-sum tax is one that does not alter the relative price ratios in the economy. A common example of a lump-sum tax is a poll tax requiring the payment of a certain amount (g in the above context) by each person in the economy.

8 Imposing a lump-sum tax of g on each young and old person would mean a lifetime tax rate of $2g$. The alternative to a tax of g on the young would be a combination of lump-sum taxes on both the young and the old in each period, with a present value in the young lifestage equal to g.

9 This is also likely to apply to many other forms of direct and indirect taxation which need laws and a bureaucracy for collection. Further, such forms of taxation impose an additional welfare loss by modifying some of the relative prices in the economy.

"shoe leather costs" and "menu costs", etc. – not captured in the preceding models. These costs are considered to be relatively insignificant for low rates of inflation, so that there is probably a net benefit from using seigniorage rather than taxation at low rates of monetary growth.

As shown above, the real value of seigniorage to the government depends upon the rate of expansion of fiat money and the amount of fiat money in real terms. This amount, for low rates of fiat money growth and inflation, is likely to be a small percentage of the fiat balances held, which in turn are only a small fraction of M1 and even less of M2. Since M1 and M2 are themselves a small fraction of GDP in most economies, seigniorage from the creation of fiat money as a percentage of GDP will usually be only a few percentage points. This amount would be more substantial at monetary growth rates sufficient to cause hyperinflation, provided that such hyperinflation does not substantially decrease the real value of fiat money holdings, as it is likely to do.

Click (1998) reported estimates of seigniorage for 90 countries for the period 1971–90. Seigniorage as a proportion of GDP ranged from 0.38 to 14.8 percent, with more than half of the countries having less than 1.7 percent and about 75 percent of them having less than 2.5 percent. Seigniorage as a proportion of government spending ranged from 1 percent to above 20 percent, with 10 countries above even this level. On average, seigniorage was about 2.5 percent of GDP and financed about 10 percent of government spending.

22.2 Fiat money and bonds in the OLG framework

Fiat money in the preceding OLG models functions merely as a store of value. This role could be equally well performed by bonds, which are defined as financial instruments paying a positive rate of interest per period. For (interest-paying) bonds to perform a store-of-value role, assume that the commodities can be *directly* exchanged against bonds, without the intermediation of money or brokerage or other transactions costs and without any lags in the transactions process. Under these assumptions, which also apply to fiat money, money and bonds would be perfect substitutes as a vehicle for saving except that bonds would have a higher rate of return. Hence, since the public would prefer the asset with the higher return, it will only use bonds, while fiat money will not be traded. The value of fiat money will be zero.

In order to create a positive demand for fiat money in the presence of interest-bearing bonds, we have to create a mechanism for fiat money to pay interest at the same rate as bonds. We now lay out one scenario for this. Assume that the *commercial banks pay interest, equal to that on bonds, on all deposits of fiat money with them*, the public deposits its fiat money holdings with the commercial banks[10] and the banks redeposit this fiat money with the central bank, which also pays the identical rate of interest on deposits with it. The banks will then function as purely pass-through financial intermediaries between the public and the central bank.[11] The return-maximizing *public will deposit all the fiat money with the banks* and not hold any fiat money (or currency, in normal terminology). Instead, the public would carry out transactions between fiat money and commodities only through the use of checks drawn on their deposit accounts, so that that there will not be any fiat money left in circulation!

10 Banks are assumed to have zero costs of servicing deposits.

11 This is similar to Wicksell's pure credit economy.

Since the bank deposits and the bonds will have the same rate of return, the public will be indifferent between them as a medium of saving, and they would be perfect substitutes. Hence, deposits (originating with fiat money) and bonds would be identical in their economic role. They would only differ superficially as accounts on the books of the banks (deposits) versus pieces of paper (bonds) held by the public. Since they are perfect substitutes, only their total quantity will be relevant for determining the price level and the values of money and bonds.

Determination of the price level when money and bonds are perfect substitutes

Note that with money and bonds as perfect substitutes, the demand for each is indeterminate and only their combined real demand can be determined. In the context of the OLG models in this chapter and the preceding one, this combined demand will equal saving. Hence, the individual's and the economy's per capita demand for money plus bonds is specified by:

$$(m^{\mathrm{d}}{}_t + b^{\mathrm{d}}{}_t)/p_t = s_t$$

where $b^{\mathrm{d}}{}_t$ is the per capita nominal value of bonds in the economy. Since $s_t = (w^{\mathrm{y}}{}_t - c^{\mathrm{y}}{}_t)$, the sum of the aggregate demand, corresponding to that from equation (21.10), for these two assets is:

$$M^{\mathrm{d}}{}_t + B^{\mathrm{d}}{}_t = N_t[p_t(w^{\mathrm{y}}{}_t - c^{\mathrm{y}}{}_t)] \tag{21}$$

The aggregate supply of money and bonds in the economy is given by their amounts held by the old (born in $t-1$, with their number as N_{t-1}) and which they want to trade. It equals:

$$M^{\mathrm{s}}{}_t + B^{\mathrm{s}}{}_t = N_{t-1}[m^{\mathrm{o}}{}_t + b^{\mathrm{o}}{}_t] \tag{22}$$

so that market clearance, with the joint demand for financial assets equal to their supply, implies that:

$$N_t[p_t(w^{\mathrm{y}}{}_t - c^{\mathrm{y}}{}_t)] = N_{t-1}[m^{\mathrm{o}}{}_t + b^{\mathrm{o}}{}_t] \tag{23}$$

Hence:

$$p_t = (M_t + B_t)/[N_t(w^{\mathrm{y}}{}_t - c^{\mathrm{y}}{}_t)] \tag{24}$$

Since $c^{\mathrm{y}}{}_t$ on the right-hand side of (24) depends on p_{t+1}/p_t, we have:

$$p_t = (M_t + B_t)/[N_t(w^{\mathrm{y}}{}_t - c^{\mathrm{y}}{}_t(p_{t+1}/p_t, w^{\mathrm{y}}{}_t, w^{\mathrm{o}}{}_{t+1}))] \tag{24$'$}$$

Hence, the current price level depends on the intertemporal price ratio p_{t+1}/p_t and varies proportionately with the *combined* supply of money and bonds, *ceteris paribus*. Therefore, an increase in the supply of bonds will raise the price level. Further, open market operations between money and bonds will not change their sum, so that such operations will not have any effects on the price level. This runs contrary to the implications of IS–LM models.

Inefficiency of changes in the supply of bonds in this model

The preceding chapter has shown the inefficiency of monetary growth. Since money and bonds are perfect substitutes in the preceding analysis of this chapter, it can be shown by a similar argument that, for a *given* quantity of money, *bond creation* will also impose a welfare loss by introducing an allocative inefficiency in the intertemporal consumption pattern. But open market operations, which maintain the sum of money and bonds but vary their proportion, are efficient.

Distinctive roles of money and bonds in real-world economies

The modern economy has distinct roles for money (whether fiat money or deposits in banks) and bonds. In daily life, money functions as a medium of payments, while bonds do not, even though each can be a vehicle for holding savings. The preceding assumptions and analyses – and the standard OLG models – do not allow this differentiation and do not allow distinctive roles for money and bonds along the ways money and bonds function in the economy. The next chapter creates this differentiation in two different ways and explores its implications in the context of modified OLG models.

22.3 Wallace–Modigliani–Miller (W–M–M) theorem on open market operations

The W–M–M theorem is an extension of the Modigliani–Miller (1958) theorem. It was first presented by Wallace (1981) and our treatment draws upon his presentation. Open market operations (OMO) are defined as the government purchase (or sale) of an asset with money, with government consumption held constant. To investigate open market operations in the OLG framework, we need to have money and another asset, both of which are simultaneously held by the public and can also be held by the government. This theorem states that *open market operations between money and another asset, with government consumption held constant, will not have any real effects (e.g. on consumption and saving) or even change the price level and nominal values of the variables*.

The intuitive explanation of the W–M–M theorem is as follows: in the assumed scenario, since the individual will be indifferent between the proportions held of money and the alternative asset (since they perform identical functions and have identical returns), policy-induced changes in their relative amounts in the hands of the public will have no consequences for the economy. We present below this theorem for two scenarios: (1) a model with commodity storage but without bonds; and (2) a model with bonds.

22.3.1 W–M–M theorem on open market operations with commodity storage

We first present this theorem for the basic two-lifestages model without bonds but modified to allow for the storage of a single commodity in the model. With this modification, both the commodity and money can act as a vehicle for saving, while bonds are, by assumption, not available in this version of the OLG model. To create simultaneously a positive demand for both the stored commodity and money, Wallace assumed that storing the commodity in period t has a stochastic real gross return ρ_t, which has a stationary distribution with $E\rho_t > 1$ and positive variance. For a distribution of ρ_t such that $E(\rho_t - 1)$ just compensates a risk-averse individual for the riskiness of the yield on the stored commodity, this individual will

be willing to hold part of his saving in money and part in the stored commodity. If he finds that the return from the commodity storage is too favorable after taking account of its risk and the degree of risk aversion in the economy, such storage will dominate fiat money and fiat money will no longer be demanded.[12] The demand for money will thus be extremely sensitive to variations in the return distribution and the degree of risk aversion. Since we need to posit positive demands by the individual simultaneously for both fiat money and the stored commodity, we will assume, though it is admittedly highly unrealistic, that the return on commodity storage is *finely* balanced at the level – and stays at this level no matter how much endowments, consumption and saving, storage or supply of fiat money change – which ensures such positive demands.

Initially, the period personal budget constraints are:

$$p_t c^{y}{}_t + m_t + p_t k_t = p_t w^{y}{}_t \tag{25}$$

$$p_{t+1} c^{o}{}_{t+1} = p_{t+1} w^{o}{}_{t+1} + m_t + p_{t+1}\, \rho_t k_t \tag{26}$$

where:

k_t = private (per capita) real holdings of the stored commodity
ρ_t = gross real rate of return on k_t.

Assume that the government purchases dk_t of the stored commodity from the young and pays in exchange an amount of money dm_t equal to $p_t dk_t$. This purchase of the commodity from the government's perspective is written as $dk^{g}{}_t (= dk_t)$. In order to keep government stocks constant at $k^{g}{}_t$, *the government transfers its net return to the old in* $t+1$ *as a lump-sum transfer* not associated with the individual's holdings of money or commodities. This net return is the growth of the commodity in storage less the loss in real value from deflation (or plus the gain from inflation) in its price and is given by $(\rho_t - p_t/p_{t+1})$. Therefore, the total amount thus received from the government in commodities equals $(\rho_t - p_t/p_{t+1}) dk^{g}{}_t$.

The government's open market purchase of dk_t of commodities against dm_t of money balances modifies the two personal budget constraints to:

$$p_t c^{y}{}_t + (m_t + dm_t) + p_t(k_t - dk_t) = p_t w^{y}{}_t \qquad m_t > 0, (k_t - dk_t) \geq 0 \tag{27}$$

$$p_{t+1} c^{o}{}_{t+1} = p_{t+1} w^{o}{}_{t+1} + (m_t + dm_t) + p_{t+1}\rho_t(k_t - dk_t) + p_{t+1}(\rho_t - p_t/p_{t+1}) dk^{g}{}_t \tag{28}$$

where $k^{g}{}_t$ are the per capita government stocks of the stored commodity, with a gross return of $k^{g}{}_t$. In (27), the young carry forward $(m_t + dm_t)$ of money and $(k_t - dk_t)$ of the stored commodity. In (28), the old have $(m_t + dm_t)$ of money. They also have $(k_t - dk_t)$ of the stored commodity plus its net return, whose total value equals $p_{t+1}\rho_t(k_t - dk_t)$. In addition, they receive a lump-sum transfer from the government of $p_{t+1}(\rho_t - p_t/p_{t+1}) dk^{g}{}_t$.

The open market operation specifies the relationship:

$$dm_t = p_t dk_t = p_t dk^{g}{}_t \tag{29}$$

Substitution of (29) into (27) and (28) yields equations identical to (25) and (26), so that there is no change in the period personal constraints and, therefore, also none in the

12 Note that for any given distribution of returns to storage, if the individuals in the economy have different degrees of risk aversion, then some will demand fiat money and others will not.

intertemporal constraint. Since the open market operations also do not change the individual's utility function, we conclude that the optimal consumption and saving paths are unchanged by open market operations.

Turning to the demand for money, from (27), the demand for real balances is given by the part of real saving that the individual wants to carry forward in the form of real balances. Let the per capita demand for and supply of nominal balances *after* the open market operations be designated respectively by $m^{d*}{}_t$ and $m^{s*}{}_t$ and those prior to the operations be shown without an asterisk. $m^{d*}{}_t$ is given by:

$$m^{d*}{}_t/p_t = s^y{}_t - (k_t - dk_t) = (s^y{}_t - k_t) + dk_t \tag{30}$$

With $(s^y{}_t - k_t) = m^d{}_t/p_t$ from (21) and with $dk_t = dm_t/p_t$ from (29), (30) becomes:

$$m^{d*}{}_t/p_t = m^d{}_t/p_t + dm_t/p_t$$

$$m^{d*}{}_t = m^d{}_t + dm_t \tag{31}$$

Therefore, money demand increases by the amount dm_t of the open market operations. Further, the open market operations also increase the money supply by dm_t, so that:

$$m^{s*}{}_t = m^s{}_t + dm_t \tag{32}$$

so that the demand for money balances in both nominal and real terms increases by exactly the same amount as their supply at the existing prices. Therefore, the price level will also not be affected by the open market operations. Nor will the nominal income, which equals endowments times the price level, be changed by these operations.

Note that in this commodity–money economy, if the money market remains in equilibrium, then, by Walras's law, the commodity market will also stay in equilibrium at an unchanged price level, even though the government in its open market operations exchanged commodities against money.

Hence, we have the W–M–M theorem that the change through open market operations in the proportions held by the individual of the two assets – money and the stored commodity – has no effect on the individual's consumption and saving paths, or on the price level. The only change is in the proportion in which the individual will hold the two assets. The change in this proportion, due to the open market operations, is willingly held by the public at the existing prices.

These results are not surprising under the given assumptions. Since both money and the stored commodity only act as abodes of purchasing power, both having an identical degree of liquidity and both having the same return net of the risk premium,[13] the young are indifferent between the actual amounts held of each at the current rate of return. The open market operations mechanism changes the amount of the commodity stored by the young but without a change in the young individual's real lifetime endowment of the commodity and without a change in his consumption and saving paths. To compensate for this reduction in his stored commodity, the rational young will want to carry forward the original amount of total saving

13 After compensation for the riskiness of the return on the stored commodity, the gross return on the stored commodity will equal p_t/p_{t+1}, which is the gross return on the riskless asset money.

by willingly holding the corresponding increase in real balances, without requiring any change in prices.[14]

Evaluating the W–M–M theorem in the context of the IS–LM analysis

Since the invariance of prices and nominal income in the W–M–M theorem is at odds with the conclusions of the macroeconomic models of Chapters 13 to 16, we illustrate its derivation in the context of the IS–LM framework. In this framework, an increase in the money supply resulting from open market operations shifts the LM curve to the right while an increase in money demand will shift it to the left. Hence, in the W–W–M analysis, with the open market operations increasing both money supply and money demand by exactly the same amounts, as shown by (31) and (32), the LM curve will not shift either way, nor will the IS and the aggregate supply curves be affected.[15] Consequently, with none of the curves of the IS–LM shifting and with commodity supply unchanged, if the economy is initially in equilibrium at the price level P_0, the open market operations will not change that price level. Nor will they change consumption and saving.

22.3.2 W–M–M theorem on open market operations in the money–bonds OLG model

To adapt the W–M–M theorem to an economy with bonds, consider an economy in which the central bank acts for the government and can store the commodity as efficiently as the public. It issues fiat money and bonds and puts these liabilities into circulation by purchasing commodities, which are stored. It earns the gross rate of return ρ on such storage, which it pays in interest on commercial banks' deposits with it and on bonds. As with the earlier analysis allowing storage of commodities, the gross return ρ is stochastic and has the same properties as in the earlier analysis. In particular, $(\rho - 1)$ just compensates for the risk of variation in the return on bonds. Note that, under these assumptions, the rate of return is the same on bonds and on banks' deposits with the central bank. The public deposits its fiat money with banks, which pay the return ρ on it, and the banks place the public's deposits with the central bank.[16] The public holds deposits and bonds and may or may not hold commodities. This scenario is an adaptation of that in the preceding subsection, except that the *central bank* now holds commodities on which it obtains a return ρ.

Since fiat money (or their equivalent deposits in banks) and bonds are equivalent in terms of being a medium of saving and pay the same rate of return, the public will be indifferent between them. The aggregate demand for money and bonds for carrying forward saving as against their aggregate supply will determine prices and the value of money. The composition of this aggregate is irrelevant in this determination. Therefore, an open market operation

14 An example from the microeconomic analysis of commodity markets can illustrate this point. Suppose an individual has bought ten red apples for his consumption. He is indifferent between red and green apples, which differ according to him only in color, which to him is a totally superficial aspect. If another person were to force or ask him to exchange some of his red apples for green ones, he would remain indifferent and take no action, such as returning to the store to buy more red apples. Without such action, there would be no change in the demand and price of red apples.

15 Note that, in the IS–LM analysis, the open market operations change the money supply but not the money demand, while in the analysis of the W–M–M theorem they change both money supply and demand.

16 Note that, under these assumptions, there is no fiat money left to circulate in the economy.

between fiat money and bonds would have no consequences for the price of commodities and for the value of money. Neither these nor the real variables of the economy will change because of the open market operations, so the W–M–M theorem will also hold for this commodities–money–bonds economy. This can be established by repeating the analysis of the previous section but replacing k by b in equations (25) through (29) – which we leave to the reader – or simply by modifying (24) to encompass market operations performed in t. These operations will increase the economy's money supply by dM_t and decrease its supply of bonds by dB_t, so that the equilibrium in the financial sector will imply the following version of (24):

$$p_t = (M_t + dM_t + B_t - dB_t)/[N_t(w^y{}_t - c^y{}_t)] \tag{33}$$

where $dM_t = dB_t$, so that:

$$p_t = (M_t + B_t)/[N_t(w^y{}_t - c^y{}_t)] \tag{34}$$

(34) is identical with (24). Hence, the open market operations do not change the price level or the value of money.

W–M–M theorem, money neutrality and IS–LM analysis

Note that the W–M–M theorem for the economy with money, bonds and/or commodity storage is stronger than the concept of the neutrality of money. The latter is that the *real* values of the variables of the economy are invariant with respect to changes in the quantity of money, but the price level and the nominal values of the variables are affected by such changes. The W–M–M theorem states that both the real and the nominal values of the variables, including the price level, are invariant to money supply changes. This result runs counter to the fundamental implications of the IS–LM analysis and many other macroeconomic models, whether neoclassical, classical or Keynesian, that an increase in the money supply through open market operations will increase nominal national income and prices, and, in Keynesian models, might also increase real output. In these models, in contrast to the W–M–M set-up for bonds, bonds are illiquid, while money is liquid, so that an increase in the money supply, with an offsetting decrease in the quantity of bonds, increases liquidity in the economy. This increase in liquidity causes increases in prices and nominal income. Therefore, the underlying assumptions on the nature of money and bonds are different between the IS–LM and the OLG models, and these differences yield the difference in the conclusions drawn from these models. Consistent with these conclusions, McCallum (1983) maintains that the distinctive implications of the OLG models arise from their lack of a medium-of-payments role for money.

Differences in the liquidity of money and bonds – and the W–M–M theorem

What is the essential nature of bonds in the real-world economy and how does it differ from that of money? Money is the only medium of exchange and the most liquid asset. Bonds, even Treasury bills, are not accepted in exchanges for commodities. While some types of bonds can be quite liquid, the medium- and long-term ones are not liquid enough to function as a medium of payments or as near-monies. If this empiricism is accepted, then bonds have to be modeled differently from what is done in the W–M–M model. In the real world, bonds pay a

positive (expected) net return and for some types of bonds, such as Treasury bills, this return is not stochastic. Such riskless assets with a positive net return dominate the non-interest-bearing money as a store of value – but not as a medium of exchange. However, bonds in the W–M–M model do not possess these properties.

As we have shown, the reason for the W–M–M theorem is that money and bonds are *perfect* substitutes under its assumptions, with both acting as a medium of payments, but the real world does not show such perfect substitution. The opposite (real world) scenario envisaging distinctive roles for money and bonds, resulting in *limited* substitution between them, is incorporated in the IS–LM models and can be incorporated into an extended OLG model augmented with cash-in-advance for purchases, or with the indirect MIUF. The applicability of the W–M–M theorem in that context is examined in the next chapter.

22.4 Getting beyond the simplistic OLG analysis of money

22.4.1 Model 1: an OLG model with money, capital and production

This section applies the OLG analysis to specific models with capital and specific utility and production functions that do not have money as an argument. Nevertheless, money is in these models because of its role as an asset in which savings can be held.

Assumptions

Assume, as in earlier sections, that individuals live through two lifestages, each lasting one period. Each individual has an endowment of one unit of labor when young and none when old. This unit of labor can be rented to firms, in return for a real wage equal to the marginal product of labor, paid in commodities at the end of the young lifestage.

Saving can be held in either money or ownership of claims to the capital of firms (i.e. shares issued by firms),[17] with capital being in the units of the commodity and equaling the stock of the commodity lent to the firms for use in further production. Money is fiat money and does not pay any interest. Capital borrowed by firms from consumers in t pays in $t+1$ gross interest r_t^{K} equal to its gross marginal product in t. The young have neither endowments of money nor capital when they are born.

The production function of the economy is:

$$X_t = AL_t^{\alpha} K_t^{\beta} \qquad \alpha + \beta = 1 \tag{35}$$

where X is output, L is the labor force and K is capital.

The utility function of each individual is:

$$U(c^{\mathrm{y}}, c^{\mathrm{o}}) = u(c^{\mathrm{y}}) + \delta u(c^{\mathrm{o}}) \tag{36}$$

where u is the period utility such that $u(c) = \ln\ c$.

Population in this economy grows at the gross rate n, with N_t persons born in period t, and the money supply M grows at the gross rate θ. m, as before, is per capita money balances purchased during the young lifestage. The increase in the money supply is distributed as a

17 Alternatively, capital can be owned by individuals, who would lend it to firms against the firms' bonds (IOUs).

gift to each of the old on a lump sum per capita basis, so that the old start the second lifestage with per capita money balances equal to θm.[18]

We consider in the following subsections a number of issues for this economy.

Prices and the growth rate of the value of money

Since the individual has an income of w_t in period t and none in $t+1$ and the increase in the money supply is distributed as a lump sum gift to the old,[19] the period constraints are:

$$p_t c^{\mathrm{y}}{}_t + m_t + p_t k_t = p_t w^{\mathrm{y}}{}_t \tag{37}$$

$$p_{t+1} c^{\mathrm{o}}{}_{t+1} = r^{\mathrm{K}}{}_t p_{t+1} k_t + m_t + (\theta - 1) M_t / N_t \tag{38}$$

where, as before, the lower-case letters designate per capita values of the variables. r^{K} is the gross return on money equal to one plus the marginal productivity of capital, and m_t is the amount of balances carried over from period t. The last term in (38) is the lump-sum receipt of newly created money. In (37), the individual can carry forward his saving in the form of either nominal balances m_t or loans to (or shares of) firms equal to capital k_t. From (37), the per capita demand for real balances is given by:

$$m^{\mathrm{d}}{}_t / p_t = w^{\mathrm{y}}{}_t - c^{\mathrm{y}}{}_t - k_t \tag{39}$$

Hence, the aggregate nominal demand for money is:

$$M^{\mathrm{d}}{}_t = p_t [N_t (w^{\mathrm{y}}{}_t - c^{\mathrm{y}}{}_t - k_t)] \tag{40}$$

The nominal supply of money in the economy is given by:

$$M^{\mathrm{s}}{}_t = M_t \tag{41}$$

From (39) and (40), equilibrium in the money market requires that:

$$M_t = p_t [N_t (w^{\mathrm{y}}{}_t - c^{\mathrm{y}}{}_t - k_t)] \tag{42}$$

which yields:

$$p_t = M_t / [N_t (w^{\mathrm{y}}{}_t - c^{\mathrm{y}}{}_t - k_t)] \tag{43}$$

If the money supply and population grow at constant rates θ and n respectively, $M_{t+i} = \theta^{\mathrm{i}} M_t$ and $N_{t+i} = n^{\mathrm{i}} N_t$. In the stationary state with these growth rates and constant values of w^{y}, c^{y} and k, (43) implies that the growth of the price level over time is given by:

$$p_{t+i} / p_t = \theta^{\mathrm{i}} / n^{\mathrm{i}} \tag{44}$$

18 Note that the gifts to the old are not in proportion to their prior holdings of money.

19 The seigniorage derived from the money creation in this section is given by:

$$\frac{\partial_{g_1}}{\partial \theta} = \left\{ -\theta \frac{m_1}{p_1} \right\} + (1 - \theta) \frac{\partial (m_1 / p_1)}{\partial \theta}$$

so that the value of money grows at the gross rate given by:

$$v_{t+i}/v_t = p_t/p_{t+i} = n^{\mathrm{i}}/\theta^{\mathrm{i}} \tag{45}$$

Interest rate and real wage in the economy

The economy was assumed above to possess the production technology:

$$X_t = AL_t{}^{\alpha}K_t{}^{\beta} \qquad \alpha + \beta = 1 \tag{46}$$

Since the net rate of return $(r^{\mathrm{K}} - 1)$ on K equals the marginal product of capital X_K, we have:

$$\begin{aligned} r^{\mathrm{K}}{}_t - 1 &= X_{Kt} = \beta AL_t{}^{\alpha}K_t{}^{\beta-1} \\ &= \beta X_t/K_t \end{aligned} \tag{47}$$

That is, if an individual lends K_t to a firm at a net rate of return X_{Kt}, he will receive $(1+X_{Kt})K_t$ in $t+1$.

With real wages equal to the marginal product of labor X_L, and the (fully employed) labor force equal to N_t, the real wage rate is specified by:

$$\begin{aligned} w_t &= X_{Lt} = \alpha\beta AN_t{}^{\alpha-1}K_t{}^{\beta} \\ &= \alpha X_t/N_t \end{aligned} \tag{48}$$

Conditions for the simultaneous positive demands for money and capital

A young individual is willing to simultaneously hold both money balances and capital if their return is identical. The gross return r^{m} on money is the increase in its value and is given by (45) as:

$$r^{\mathrm{m}}{}_t = v_{t+1}/v_t = p_t/p_{t+1} = n/\theta$$

Both money and capital will be held if:

$$r^{\mathrm{m}}{}_t = r^{\mathrm{K}}{}_t$$

which requires that:

$$n/\theta = 1 + \beta X_t/K_t \tag{49}$$

This yields the required rate of price increase as:

$$p_{t+1}/p_t = [1 + \beta X_t/K_t]^{-1} \tag{50}$$

Capital and output in the economy

Given θ and n, (49) specifies the amount of capital held in the economy as:

$$K_t = (n/\theta - 1)^{-1}\beta X_t \tag{51}$$[20]

so that the representative individual's demand for and holdings of capital, equal to the per capita amount of capital held by the young, are:

$$k_t = K_t/N_t = (n/\theta - 1)^{-1}\beta x_t \tag{52}$$

where $x = X/N$, so that the representative individual in t will carry forward nominal balances of m_t and real capital of k_t. Note that the amount of capital increases with the rate of money growth (which makes carrying forward money balances less attractive) and decreases with the rate of population growth (which makes carrying money balances more attractive).

With K_t specified by (51), output produced in the economy is given by (35) and (51) as:

$$\begin{aligned} X_t &= AN_t^{\alpha}\{(n/\theta - 1)^{-1}\beta X_t)\}^{\beta} \\ &= [AN_t^{\alpha}\{(n/\theta - 1)^{-1}\beta)\}^{\beta}]^{1/(1-\beta)} \end{aligned} \tag{53}$$

where $\partial X_t/\partial\theta > 0$ and $\partial X_t/\partial n < 0$, so that output increases as the population growth rate decreases and the money growth increases.[21] Hence, money growth is not neutral and has a positive impact on capital and output by making it less attractive to hold money balances. Clearly, the W–M–M theorem does not hold in this model.

Pareto dominance of the trade equilibrium with money over autarchy

For the preceding model, the period personal budget constraints were:

$$p_t c^{y}{}_t + m_t + p_t k_t = p_t w^{y}{}_t \tag{54}$$

$$p_{t+1} c^{o}{}_{t+1} = r^{K}{}_t p_{t+1} k_t + m_t + (\theta - 1)M_t/N_t \tag{55}$$

Combining these equations by eliminating the money balances carried forward gives the personal lifetime budget constraint as:

$$c^{y}{}_t + (p_{t+1}/p_t)c^{o}{}_{t+1} = w^{y}{}_t + \{(p_{t+1}/p_t)r^{K}{}_t - 1\}k_t + (\theta - 1)m_t/p_t \tag{56}$$

Since $p_{t+1}/p_t = \theta/n$ from (44), (56) becomes:

$$c^{y}{}_t + (\theta/n)c^{o}{}_{t+1} = w^{y}{}_t + \{(\theta/n)r^{K}{}_t - 1\}k_t + (\theta - 1)m_t/p_t \tag{57}$$

20 For this relationship, $\partial K_t/\partial\theta > 0$ and $\partial K_t/\partial n < 0$. The reason for the latter is as follows. As n increases, endowments rise at a faster rate, which increases the rate of deflation. This causes the return on money to rise, so that, out of a given saving, more money is held while the amount held as capital decreases.

21 This is rather a strange result, whose explanation is as follows. As population grows, it causes a decline in price level, so that the value of money rises. Hence, out of a given saving, more money and less capital is held. With less capital, the production function yields lower output.

For this model, the utility function was assumed to be:

$$U(c^{y}, c^{o}) = u(c^{y}) + \delta u(c^{o}) \tag{58}$$

In the case where money is used, maximizing the utility function (58) subject to the budget constraint (57) yields the optimal consumption pattern as:

$$c^{o}{}_{t+1} = (\delta n/\theta) c^{y}{}_{t} \tag{59}$$

and

$$c^{y}{}_{t} = (1+\delta)^{-1} W_t \tag{60}$$

where W_t equals the right-hand side of (57). Let the derived optimal consumption pattern be $c^{y}{}_{t}{}^{*}$ and $c^{o}{}_{t+1}{}^{*}$. Substitution of these in the utility function shows that these values will yield higher utility than the autarchic solution and will be Pareto optimal.

22.4.2 Model II: the preceding OLG model with a linear production function

If, instead of a Cobb–Douglas technology specified by (35), we suppose that the production technology of the economy had been a linear one of the form:

$$X_t = aL_t + bK_t \tag{61}$$

In this case, we would have $(r^{K}{}_{t} - 1) = X_{Kt} = b$, so that the condition for the equality of the return on money and capital would be $n/\theta = b$. It can be shown that if this condition is satisfied, neither the stock of capital nor the output of the economy is uniquely determined. However, assuming the independent determination of θ, n and b, it would indeed be fortuitous if this condition were satisfied for any economy, let alone for most economies. The more likely possibility is that it will not be met. If it is not, and if $(n/\theta - 1) > b$, only money will be held. But if $(n/\theta - 1) < b$, only capital will be held.

Hence, the model with a linear production function also shows the money growth rate to be non-neutral, since for some values of θ the equilibrium amount of capital will be zero, while it would be positive for other (higher) values of θ. Hence, output would be higher under the latter than the former condition. The W–M–M theorem therefore does not hold in this model with a linear production function – and did not hold in the earlier model with a Cobb–Douglas technology.

22.5 Model III: the Lucas OLG model with non-neutrality of money[22]

The specific OLG models of the preceding section with a production technology showed that money need not be neutral in the OLG models. In these models, while a production technology was specified and capital was variable, the supply of labor was kept exogenous.

22 This section is based on Lucas (1996).

This section investigates this issue for a structure of the economy that allows the supply of labor to vary with its wage rate.[23]

Assume, though still within the two-lifestages OLG framework, a constant population and that each young person supplies h units of labor but does not consume in the young lifestage, while, when old, he consumes but has no labor to supply. The labor supply and consumption pair is $(h^{\mathrm{y}}, c^{\mathrm{o}})$. Let the lifetime preference function be:

$$U(h^{\mathrm{y}}, c^{\mathrm{o}}) = -h^{\mathrm{y}} + u(c^{\mathrm{o}}) \tag{62}$$

where labor/work has been assigned a negative utility and $u(\cdot)$ is the period utility function. Further, assume that the production technology is a simple one such that one unit of labor yields one unit of the consumption good. With zero consumption in the young lifestage, output in the young lifestage is converted into money and carried to the following period.

As in the earlier sections of this chapter, assume that the money supply grows at a constant gross rate θ and the new money is distributed to the old as a lump-sum transfer.

The young supply h units of labor and receive h units of the commodity, but do not have any consumption. The old do not work but receive a lump-sum gratuitous transfer of the newly created money. Under these assumptions, the lifetime budget constraint is given by:

$$c^{\mathrm{o}}{}_{t+1} = \{p_t h_t + (\theta - 1)m_t\}/p_{t+1} \tag{63}$$

Substituting this in the individual's lifetime utility function (62) yields:

$$U(h_t, c^{\mathrm{o}}{}_{t+1}) = u\left(\frac{p_t h_t + (\theta - 1)m_t}{p_{t+1}}\right) - h_t \tag{64}$$

Maximizing (64) with respect to h_t gives the first-order condition as:

$$u'(\cdot)(p_t/p_{t+1}) - 1 = 0$$

which implies that:

$$u'\left(\frac{p_t h_t + (\theta - 1)m_t}{p_{t+1}}\right) = \frac{p_{t+1}}{p_t} \tag{65}$$

Assume that the solution to this equation is $h^*{}_t$. In the stationary state with $h^*{}_t$ constant at h^* and a constant population, output will be stationary while the money supply will be increasing at the gross rate θ. Consequently, prices will also increase at the gross rate θ, so that (65) implies that:

$$u'(\cdot) = \theta \tag{66}$$

Equation (66) requires that an increase in θ be met by an increase in marginal utility, which requires a decrease in h. Hence, h^* – and, therefore, production and saving – will be a

23 The variation in models is being deliberately offered in an attempt to increase exposure to the types of models that can be used within the OLG framework.

decreasing function of the rate of inflation and monetary growth, with the result that the money growth rate is again non-neutral, as in the preceding section. However, while increases in the money growth rate increased output in the models of the preceding two sections, they decrease output in this section. The reason for the latter is that the faster rate of money growth decreases the return to saving, which equals the income of the young from their labor supply, so that the faster money growth rate ends up increasing the inflation-based tax on labor income and reduces the labor supply.

An adaptation of the preceding model

The inflation-based tax on labor supply would be eliminated if the increase in the money supply were to be distributed not as a lump sum to the old, but in proportion to money holdings, which equal $p_t h_t$. In this case, the lifetime budget constraint becomes:

$$\begin{aligned} c^{o}{}_{t+1} &= \{p_t h_t + (\theta - 1)p_t h_t\}/p_{t+1} \\ &= h_t(\theta p_t/p_{t+1}) \end{aligned} \tag{67}$$

With $(\theta p_t/p_{t+1}) = 1$ in the stationary state with a constant population, (67) becomes:

$$c^{o}{}_{t+1} = h_t \tag{68}$$

so that the individual's lifetime utility function becomes:

$$U(h^{y}{}_t, c^{o}{}_{t+1}) = -h_t + u(h_t) \tag{69}$$

Maximizing (69) with respect to h_t implies that:

$$u'(h_t) = 1 \tag{70}$$

so that the supply of labor, as well as output and saving, is independent of the rate of money growth, making money neutral – in comparison with (65) which had implied its non-neutrality, with a negative impact of monetary growth on output growth. Hence, the manner in which the seigniorage from the monetary creation is distributed matters for the neutrality of money.

Lucas (1996) showed that further adaptations of the above model could also generate positive transitional effects of monetary growth. An example of the latter occurs if there are expectational errors of relative price changes, as in some of Lucas's other contributions (see Chapter 14).

Note that both the neutrality of money (which allows the impact of changes in the money supply on prices and nominal variables but not on output and employment) and the non-neutrality of money (which allows the impact of changes in the money supply on prices and nominal variables, as well as on output and employment) are at odds with the earlier W–M–M theorem, which asserts that there would not be any impact, *even on prices or nominal variables,* of increases in the money supply. However, this theorem, unlike the three preceding models, assumes the existence of bonds (or a stored commodity) that are perfect substitutes for money and that the increases in the money supply occur through open market transfers, so that each individual's wealth remains unchanged by open market operations. The precise assumptions of the model and the mechanism for changes in the money supply are

therefore critical to conclusions on the impact of money on the economy and to its neutrality. To conclude, the OLG framework on its own does not imply either the W–M–M theorem or the neutrality of money; these come from other assumptions of the model.

22.6 Do the OLG models explain the major facets of a monetary economy?

The major doubt or objection about the validity of the OLG framework as a way of modeling money in the monetary economy concerns the role of money as a medium of exchange. The fiat money in the standard versions of the OLG models clearly serves as a store of value or a temporary abode of purchasing power. It trades against commodities and enables the multilateral exchange of commodities between members of different generations over time, so that its use increases allocative efficiency in the consumption of commodities over time. These properties seem to be aspects of a medium-of-exchange role and many proponents of the OLG models (e.g. Wallace, 1980, p. 77) have argued that money does function as a medium of payments in these models since its use is the only way the young and the old can trade in the simple OLG framework.[24] Further, the proponents of the OLG models argue that money does not properly belong in the utility function and such functions should not be introduced directly or indirectly into more elaborate forms of the OLG models.[25] The OLG paradigm is then proposed as an alternative to models with money in the utility function (MIUF) or the production function (MIPF), and to those with a cash-in-advance constraint, and is claimed by some of its proponents to be the best of the available paradigms for representing money in the modern economy (Wallace, 1980, p. 50).

Medium of exchange/payments role of money

The critics of the OLG framework respond that while money does act as a store of value in it, it does not really act as a medium of exchange in those transactions in which saving is not an integral part of the transaction as intended by either party, and it is this role which is essential to the nature of money in the modern economy.[26] Intuitively, in the OLG model, there is no role for money in exchanges among the young or among the old, or in payments made by the old to the young, while these are ever-present in a monetary economy. Money's other role in facilitating multilateral exchanges across generations could be wholly performed by other assets (e.g. bonds) that are clearly not media of exchange. To illustrate, if the return on the storage of commodities were higher than on fiat money, the latter would not be traded and the postponement of consumption from the young to the old lifestage would be accomplished

24 Note that such a role is only a very limited part of the transactions role of money in the real world, where most of the transactions between commodities and money are for current exchanges and do not involve any intended saving.

25 For example, Wallace (1980, p. 49) is critical of the money-in-the-utility-function approach, arguing that doing so "begs too many questions [about which money appears in the utility function and why]… [and] theories that abandon intrinsic uselessness will be almost devoid of implications."

26 Tobin (1980, p. 83) claims that "The 'consumption–loan' parable is valuable and instructive, but it should not be taken seriously as an explanation of the existence of money in human society…. One can call the fiat store of value of the model *money*, but it bears little resemblance to the money of common parlance or the money that economists and policy makers argue about." Among his many other criticisms of the OLG models is "isn't it ridiculous to identify as money the asset that the typical agent would hold for an average of 25 years, say, from age 40 to age 65? The average holding period of a dollar of demand deposits is about 2 days" (p. 84).

directly by commodities, which clearly would not be acting as a medium of exchange. Since the storage of commodities has been an option available in all economies and its return would have fluctuated with the commodity stored and the economic environment, the OLG models predict that the demand for fiat money would have fluctuated between zero and a positive amount in economies over time. Further, if the return on the storage of commodities varied sufficiently across individuals in an economy, depending upon their storage technologies, fiat money would be held by some and not by others. In particular, OLG models imply that money will not be held by dis-savers, among whom belongs the old generation. These predictions run counter to the observation that money holding is consistently a universal phenomenon across individuals and over time in modern economies.

Simultaneous existence of money and bonds with different rates of return

The modern economy has many types of bonds with differing rates of return. Some of them (such as money market instruments) definitely possess a higher rate of return net of the risk premium than fiat money, yet fiat money continues to be used in such an economy. However, in OLG models, if bonds have a higher rate of return than money, fiat money will not be used. Hence, the OLG models fail to satisfactorily explain the demand for money in real-world conditions with a positive expected return on bonds.

Note also that, in OLG models, the actual usage of money as against that of bonds and stored commodities does not provide any additional benefits other than the allocative efficiency made possible by the latter two assets. This runs counter to the generally recognized benefits of a monetary over a barter economy. The most important of these is that a monetary economy with given labor and capital inputs produces a substantially larger output of commodities than a barter economy, *ceteris paribus*, in addition to the allocative efficiency of intertemporal exchanges done through the intermediation of bonds. The use of money in current exchanges – as distinct from intertemporal exchanges over time – "matters" vitally in a qualitative sense, providing a substantial increase in both output and social welfare in a monetary economy over that in the barter economy through a better allocation of output *and* a greater output.[27] Chapter 24 on money in long-run growth will investigate the sources of the contribution of money and the financial sector to the growth of output.

The continued use of money in hyperinflations

Another well-established fact of monetary economies is the tenacity of the continued use of money under conditions in which there are persistent and high rates of increases in the money supply, even under conditions in which (some types of) commodities can be easily stored and have a net increase in value over time, and even when there are also bonds with positive real rates of interest. The standard OLG models imply that the return on fiat money would be negative under such conditions and money would be inferior to commodities and bonds, so that it would not be demanded and used in such cases. This implication is clearly counterfactual. In fact, money continues to be traded even in hyperinflations, even though it is known to be rapidly losing its purchasing power. In other words, the use of money in the

27 To allow the use of money to create a greater output, money balances must be an element of the production function or their usage must allow a saving of other inputs. The proponents of OLG models tend to shun both of these possibilities. They are considered in the next chapter in the context of a hybrid OLG model.

economy is not tenuous, as the OLG models imply, nor do economies make periodic switches from monetary to non-monetary economies – autarchic, barter, or exchange economies with bonds but without non-interest-paying money – for the types of cases implied by the standard OLG models.

Does monetary expansion decrease welfare, even with price stability?

A central implication of the OLG models is the inefficiency of monetary expansion relative to that of monetary stability, even when the monetary expansion ensures price stability. This runs counter to the purported belief of many central banks in Western countries that price stability or a low rate of inflation is preferable – in terms of increasing social welfare – to monetary stability. Further, as discussed in Chapters 11 and 12 on the central bank, central banks and economists generally believe that the ultimate objective of monetary policy should be formulated in terms of the desirable rate of inflation rather than of the rate of increase of the monetary aggregates, with the latter as intermediate targets in achieving the pre-determined ultimate goals.

Money, velocity, near-monies and innovations in the payments system

Monetary economies consistently show a velocity of circulation of fiat money greater than one. Among the reasons for this is the existence of near-monies, which are created privately and whose creation cannot be rigidly controlled by the state. Frequent innovation, resulting in the creation of close substitutes for fiat money or otherwise economizing on the use of such money, is a common phenomenon. OLG models do not seem to be able to handle this process and impose rigid limitations on the role of the private creation of near-monies and innovations in the payments process, as well as on the evolution of velocity.

The OLG models with money, the short run IS–LM model and business cycles

The money market of the monetary economy is handled in the standard IS–LM models through the concept of the LM curve, with changes in the real money supply due to open market operations (or price level changes) shifting this curve. *Wallace's Modigliani–Miller theorem implies that such operations do not shift the LM curve and have no impact on aggregate demand, the price level, the interest rate or output*. Neither the empirical studies on this topic nor the established wisdom, as represented by the treatment of the LM curve and the views of central banks on this issue, support this theorem.

Many monetary models, but especially OLG ones, assume full employment. However, this is not the universal state of all monetary economies, which go through recessions and booms. Therefore, for the modern macroeconomy, a serious monetary model must also be able to present analyses of the *disequilibrium* states – with deviations from full employment in the form of involuntary unemployment or employment above the long-run equilibrium level – and the impact of alterations in the money supply on unemployment. In other words, any reasonably adequate monetary model must be able to deal in a realistic manner with recessions and booms, as well as the consequences of money supply initiatives in such states. The current offerings of the OLG models do not do so. It remains to be seen how they will be extended to disequilibrium states and whether their implications will be consistent with the very considerable amount of data and empirical findings on the subject. However, there

does seem to be some room for skepticism about whether they will achieve this consistency any time soon.

In defense of the OLG models

It should be recognized that the OLG models with money have provided an interesting alternative to the existing MIUF and cash-in-advance paradigms and have shown new insights into the role of money in the economy. Further, there seems to be considerable room for extensions and modifications. If we do not have to be purists, an attractive possibility would be to introduce elements of the medium-of-exchange role of money (as in money directly or indirectly in the utility function or in the cash-in-advance models) into the OLG models. This is done in the next chapter.

Conclusions

In the stationary state of the OLG models with positive saving in the first lifestage, money has a positive demand, as long as its rate of return is not dominated by the return on other stores of value. To create simultaneous demand for money and other assets, the ways in which other assets are introduced into the OLG model are sometimes such that the individual becomes indifferent between holding fiat money or the other assets – whether they be the storage of the commodity or bonds. Consequently, for certain sets of assumptions on the structure of the economy and the mode of injecting money supply increases into it, open market operations between such assets have no effects on the economy and lead to the Wallace–Modigliani–Miller theorem of open market operations. This leads to the surprising implication that the monetarization of the very considerable public debt of many countries will not change the price level in those countries. This violates the stylized facts and runs counter to one of the most commonly held beliefs in monetary economics.

This chapter has also presented other OLG models in which increases in the money supply do increase prices and nominal income, with such increases being neutral or non-neutral in their impact on real output and other real variables. However, even these models still suffer from the assumption that money is no more and no less liquid than the other assets, and that the only role performed by money is to carry saving in whole or in part to the future.

In the developed monetary economy, current saving during say a month or a quarter, let alone a longer period, and accumulated wealth are rarely held totally in the form of money balances. In the absence of transactions costs, most of one's wealth is converted to other financial assets paying higher rates of return. The standard OLG models do not properly capture such patterns of money and bond holdings, even though these patterns are central to the modern financial systems. They do not do a satisfactory job of explaining the usage and effects of money as a medium of exchange in the monetary economy, and many of their distinctive implications flow from this neglect (McCallum, 1983).

From a wider perspective, fiat money is a social contrivance, for which there are usually alternative social or public contrivances that can substitute for its role of smoothing consumption over time.[28] A traditional arrangement of this type was the extended family system in which the old members of the family were supported by the young ones.

28 The most important private contrivance for transferring purchasing power through commodities from the working to the retired lifestage is home ownership.

Currently common public contrivances across generations are social security arrangements, such as government pension plans, to which individuals make tax payments during their working years and pensions are given to them by the government in their retirement. In a more general context, governmental schemes (such as social security schemes, national health care systems, unemployment insurance and welfare benefits) reduce savings for retirement or/and precautionary purposes. In OLG models, they reduce the demand for money by reducing savings. However, in reality, while such schemes reduce saving, they do not equally reduce the demand for money.

Economists are divided in their faith in the pure OLG models as *positive* theories capable of explaining and predicting the demand for money and the price level. Wallace (1980) claimed that the OLG model of fiat money includes what is essential for a good theory of money and that it "gives rise to the best available model of fiat money" (p. 50). However, many other economists reject it as being unsuitable for modeling money in the modern economy. McCallum (1983) claims that many of the distinctive implications of OLG models flow from the absence in them of a medium-of-payments role for money, which constitutes a fatal flaw for an analysis of a monetary economy. Their rejection has led to the claim that the overlapping generations framework "should not be taken seriously as an explanation of the existence of money in human society" (Tobin, 1980, p. 83).

The implications of the standard OLG theory are refuted by a great deal of empirical evidence on the stylized facts about money in the economy listed at the beginning of this chapter. These refutations are to: its money demand function, with money demand a function of saving rather than of total income or expenditure; a velocity of circulation of one per period; no impact of open market operations on the price level, on interest rates, on nominal national income, or on real output and unemployment even in disequilibrium, etc.

Summary of critical conclusions

- Seigniorage as a taxation device is inferior to lump-sum taxes, assuming no costs of revenue collection.
- If money and bonds are perfect substitutes, the price level will depend on their sum and not merely on money supply. Further, the Wallace–Modigliani–Miller theorem will apply to open market operations between them, so that changes in the money supply will not change the price level.
- Depending on the specifications of the model, money neutrality need not hold in OLG models of money.
- Many of the implications of the OLG models of fiat money are rejected by the empirical evidence. In particular, the most significant scale variable for the demand for money is not saving, but rather expenditure on commodities.

Review and discussion questions

1. What does M1 do in the real-world economies? What do bonds do? Why are they not perfect substitutes for the ordinary individual? Discuss some of the ways in which the OLG models with money might be modified to introduce bonds that are not perfect substitutes for M1 and capture the roles that you have attributed to M1 and bonds.

2. Compare the relationships between money, saving and wealth in the asic OLG model with money with those in the IS–LM and AD–AS models.
3. What is the impact in the OLG model with money of competitively created substitutes for fiat money by private commercial banks? What is the impact of innovations in banking technology, such as ATMs?
4. For the production and utility functions specified in Section 22.4 and zero population (equal to labor force) growth, examine the consequences of serious deflation for the average productivity of labor, employment and output. Is money neutral?
5. Given the information in the preceding question related to the model in Section 22.4, would this economy always maintain full employment (at a positive real wage)? If not, derive the rate of unemployment as a function of the money growth rate.
6. Compare the assumptions and implications of the OLG models of this chapter with the IS–LM and AS–AD models for (a) exogenously given endowments of commodities and without labor explicitly in the analysis; (b) the labor market and production functions of the type usually assumed in the IS–LM and AD–AS models.
7. Assuming full employment, does the Wallace–Modigliani–Miller theorem hold in the IS–LM and AD–AS models? Discuss.

References

Click, R.W. "Seigniorage in a cross-section of countries." *Journal of Money, Credit and Banking*, 30 1998, pp. 154–71.

Lucas, Robert E. "Nobel lecture: monetary neutrality." *Journal of Political Economy*, 104, 1996, pp. 661–82.

MaCallum, B.T. "The role of overlapping generations models in monetary economics." *Carnegie-Rochester Series on Public Policy*, 18, 1983, pp. 9–44.

Modigliani, F., and Miller, M.H. "The cost of capital, corporate finance, and the theory of investment." *American Economic Review*, 48, 1958, pp. 261–97.

Tobin, J. "Discussion." In J.H. Karekan, and N. Wallace, eds, *Models of Monetary Economics*. Federal Reserve Bank of Minneapolis, 1980, pp. 83–90.

Wallace, N. "The overlapping generations model of fiat money." In J.H. Karekan, and N. Wallace, eds, *Models of Monetary Economics*. Federal Reserve Bank of Minneapolis, 1980, pp. 49–82.

Wallace, N. "A Modigliani–Miller theorem for open market operations." *American Economic Review*, 71, 1981, pp. 267–74.

23 The OLG model of money

Making it more realistic

This chapter represents modifications and deviations away from the benchmark OLG models of money in Chapter 21, even though the overall framework is still that of overlapping generations. This is done to make the OLG modeling of money more realistic in replicating the way money actually functions in the economy, and to derive implications that are closer to the empirical findings on the money demand function. These modifications also allow a distinction between money and bonds; bonds pay a higher return than money but are less liquid.

Each generation is now assumed to live for T periods and in each period overlap with $(T-1)$ other generations.

In one of the modifications of the OLG model, cash-in-advance is required to pay for all purchases of commodities and bonds. In the second modification, money saves on transactions time in making purchases, so that it enters the utility and production functions in an indirect manner.

Key concepts introduced in this chapter

- T-period ($T > 2$) lifetimes
- Cash-in-advance for transactions
- Liquidity preference and the term structure of interest rates in the T-period OLG model
- The Wallace–Modigliani–Miller theorem for open market operations when money and bonds differ in liquidity
- Money in the indirect utility function
- Money in the indirect production function

The preceding two chapters presented the basic OLG model with money, and its analysis. Many of its implications were implausible for the modern economy with well-developed financial systems: they were inconsistent with the stylized facts on money presented at the beginning of Chapter 21. Its distinctive implications could arise from the use of a two-period horizon or its neglect of a medium-of-payments role for money (McCallum, 1983). This chapter explores whether extending the OLG model from 2 to T periods or providing a medium-of-payments role for money (through blending it with the cash-in-advance concept or with the money in the utility function (MIUF) and

money in the production function (MIPF) concepts) would make more of its implication valid.[1]

We now add the following to the nine stylized empirical facts listed in Chapter 21.

10. For a given real wage and population, the use of money by workers increases their labor supply for commodity production.
11. The use of money by firms increases their output of commodities, even for a given employment of labor.

We also want to reiterate at this stage a fundamental aspect of the monetary economy: as pointed out in Chapter 21, in a monetary economy, commodities, bonds and labor exchange against money but not directly against other commodities, bonds and labor. For monetary analysis, this empirical reality constitutes the *environment of the modern economy*, so that any analysis of the modern economy has to be based within this environment.

The OLG model arose as a contender to the traditional MIUF–MIPF approach to the role of money in the economy. It would therefore seem appropriate to first blend the OLG and cash-in-advance models to determine whether their implications would offer an attractive alternative in terms of empirical applicability to the MIUF models. This blend, in the context of a T-period model, is presented in Section 23.1. It is shown that it still has several implausible implications, e.g. for the close relationship between saving and the demand for money.

As discussed in Chapter 3, one justification for including money in an *indirect* utility function is that money balances reduce time in making payments in a monetary economy. In such a context, while the *direct* utility function would have only consumption and leisure in it, the amount of leisure is reduced by the time spent in carrying out transactions; so that holding money increases leisure and therefore indirectly yields positive utility. Such justifications for the MIUF emphasize that in a monetary economy it is the *environment* that allows the individual to directly or indirectly attach utility to money balances. Of the direct and indirect versions of the MIUF approach, the payments/transactions time version seems to do less violence to the origins of the OLG model. It is, therefore, the version adopted in this chapter for blending with the OLG model. This is done in Section 23.2.

As Chapter 3 showed, money can also be introduced directly or indirectly in the production function (MIPF) approach. This chapter uses the indirect MIPF approach, in which holding adequate money balances for payments and receipts reduces the usage of labor and capital in payments/transactions. The specification of the indirect MIPF approach is given in Section 23.3. Section 23.4 integrates the indirect MIUF and MIPF approaches into the basic two-lifestages version of the OLG model.

Wallace's Modigliani–Miller (W–M–M) theorem on open market operations between money and bonds or between money and a stored commodity ceases to hold in OLG models with cash-in-advance features, or those with direct and indirect MIUF and MIPF features. In particular, open market operations can affect the price level. Further, the demand for real balances ceases to be closely related to saving. Hence, the implications of the modified models are consistent with the stylized facts on money.

Most of the symbols used in this chapter are carried over from the preceding ones on the OLG models and are defined there. Only the new ones are defined in this chapter.

1 Brock (1990) provides a good treatment of the relevant issues.

23.1 A *T*-period cash-in-advance money–bonds model

In order to create a positive demand and positive value for money in the presence of bonds that yield a higher return, we need to add assumptions that allow fiat money to possess a correspondingly higher liquidity than bonds. We sketch below one possible set of such assumptions.

Assume that:

(i) The individual's lifetime is now divided into $T + 1 (T > 0)$ lifestages from 0 to T.
(ii) Purchases of either commodities or bonds require the prior possession of money.
(iii) Bonds have a minimum maturity of one period and cannot be cashed before maturity.
(iv) There is no resale market in bonds.
(v) Commodities and bonds trade only against money, but not against each other.
(vi) *Only one transaction can be conducted between commodities and money, or between bonds and money, in a period*.[2]

Distinction between money and bonds

The requirement that individuals must pay in money (labeled in this part of the literature as "cash," though this term is meant to include bank deposits) for their purchases of commodities and bonds is known as the *cash-in-advance* or *Clower constraint*, and is a fundamental requirement of a monetary economy. As pointed out earlier, it is part of the environment of the modern economy. The above assumptions imply that money is more liquid, in the sense of being useable for purchasing commodities *in* the period after its acquisition, while bonds cannot be so used; bonds cannot be cashed until the period after their acquisition and the funds thus obtained can only be used to buy commodities one period later. Intuitively, in the model being set up, with the very short durations of periods, the *prior* holdings of money can be used to buy commodities (and bonds) without further delay while the *prior* holdings of bonds require a delay of at least one period in purchasing commodities; bonds converted into money in the current period can only be used to purchase commodities in the next period. Hence, money is more liquid than bonds, so that the preceding assumptions serve our purpose of introducing a liquidity difference between money and bonds in the economy.

Duration of T *periods and the magnitudes of the variables*

Each individual in this extended model has been assumed to live for $T + 1$ periods from t to $t + T$ (inclusive). If a lifetime is taken to be 70 years, $T + 1$ would be 70 if each period equals a year; $T + 1$ would be 280 if each period is a quarter, and so on. By allowing T to become increasingly larger, the chronological equivalent of a period can be made very short. A given individual's lifetime would overlap with the lifetimes of T generations other than his own. The magnitudes of births, consumption, saving, interest rates and other flow variables would depend upon the duration of each period, each being smaller the shorter the period.

2 That is, commodities cannot be exchanged against bonds, or vice versa, in the same period but both can be exchanged against prior money holdings.

Saving, money and bonds in the extended model

It was a fundamental characteristic of the two-lifestages OLG model that there be (positive) saving in the young lifestage and dissaving in the old lifestage, with money being held as the vehicle for saving from the former to the latter. The extended model makes the corresponding assumption that there be saving and positive wealth accumulation in the early years of life, with dissaving in the latter years.[3] In the absence of bequests, the wealth remaining in the final period $t + T$ is spent on commodities in that period.

In this extended model, money is assumed to be the only medium of exchange and payments. Bonds and commodities are excluded from this function. That is, money is the medium of exchange and payments between commodities of different periods, between bonds of different periods and between commodities and bonds. Hence, (positive) saving out of the commodity endowments must be first exchanged for money, which is held for one period and then exchanged for bonds. Similarly, when bonds mature, bond holders receive money, not commodities, with the money thus received held for one period.

We have assumed that only one type of transaction can be performed in each period. That is, if money is received from the sale of either bonds or the commodity in a given period t, it cannot be exchanged for either commodities or bonds until the next period $(t + 1)$. If one-period bonds are then (in $t + 1$) bought with money, they would have to be held for the period $(t + 1)$ and can be cashed in period $(t + 2)$. The cash thus obtained in $(t + 2)$ cannot be used to pay for commodities or other bonds in that period but can be used for their purchase in $(t + 3)$. Hence, while money can be used to buy either commodities or bonds in every period, bonds and commodities cannot, so that money is the only perfectly liquid (i.e. fungible within a given period) good in the model.

Term structure of bonds, the yields to maturity and liquidity preference

There can be a variety of private or public bonds, each differing in the term to maturity from other bonds (and money), subject to the condition that the bonds must mature during the lifetime of the issuer. Since the minimum maturity is one period and $t = 0$, $1,\ldots,T$, there can be $T - 1$ types of bonds, each with a different maturity and its own coupon rate. Assuming Hicksian liquidity preference, as explained in Chapter 20, for shorter maturity by lenders and longer maturity by borrowers as in the expectations theory (incorporating liquidity preference) of the term structure of interest rates, there would be $T - 1$ coupon rates increasing in the term to maturity. Fiat money would have a zero coupon.

Starting with period 1, assume that the individual has positive saving in period 1 and negative saving in periods 2, 3 and 4.[4] He wishes to consume the period 1 saving in periods 2, 3 and 4. To finance the latter under our assumptions, he must hold the period 2 and 3 dissaving in money from period 1 onwards; if he wants to invest in one-period bonds, he can purchase them only in period 2, hold them from period 2 to 3, cash them in period 3 and use the money for dissaving – that is, buying commodities – in period 4. Hence, a τ-year bond requires holding money equal to its initial value for one period and its maturity value for one period, and cannot be used for the postponement of consumption by less than $(\tau + 3)$ periods.

3 This is really the statement that consumption must be different from the receipt of endowments in at least two periods, an earlier one with positive saving and a later one with negative saving.

4 Period 1 is for $i = 0$, period 2 is for $i = 1$, and so on.

This follows from the assumption that in a monetary economy commodities exchange against money, not against bonds, and that bonds exchange against money, not commodities, and that there is only one exchange per period.

If the coupon rate is increasing in the term to maturity, a saver saving for τ periods – that is, with a net saving in a period $(t+i)$ followed by a corresponding amount of dissaving in period $(t+i+\tau+1)$ – would prefer to buy $\tau-2$ period bonds. He would not buy a sequence of shorter term bonds since these would give him a smaller total return – with "fallow" periods in which the non-interest-bearing money has to be held in going from one bond to another – so that even if the market short rate was constant (that is, the yield curve was horizontal), the net return per period to the individual saving for τ periods would be made higher by buying $\tau-2$ period bonds than from any combination of shorter term bonds.[5] Hence, the effective rate of return for the individual would be increasing in the term to maturity even when the coupon payment per period is identical, and the markets in bonds would be segmented in the term to maturity. The effective yield curve would be upward sloping and concave.

23.1.1 Cash-in-advance models with money and one-period bonds

While the preceding assumptions allow for the existence of a multitude of bonds differing in maturity in the OLG model, we will henceforth simplify by assuming that there exist only one period bonds and they can be rolled over instantly – that is, without a delay of one period which would occur if they were first converted into money – into new bonds (but not into commodities). For the initial period t, the individual will not be able to finance dissaving in $(t+1)$ or $(t+2)$ by holding bonds: his saving of commodities in period t is exchanged for money in t; if it is converted from money into bonds in $(t+1)$, it would mature in $(t+2)$ and be repaid in money in $(t+2)$, which allows it to be spent on commodities in $(t+3)$ but not in $(t+2)$. Clearly, the OLG model with such a pattern must have at least four periods.

23.1.2 Analysis of the extended multi-period OLG cash-in-advance money–bonds model

Assume that the individual lives for $T+1$ periods from $t=0,1,\ldots,T$, so that he has $T+1$ lifestages and receives some commodities in each lifestage until his retirement. This endowment path is assumed to be concave as in the life cycle consumption theories, with income increasing in the early lifestages but eventually declining. There are zero endowments in the retirement lifestages. Peak "income" from endowments occurs at some time prior to retirement. The essential assumptions are that saving be positive in the early lifestages and negative in the retirement lifestages, and that the accumulated wealth be positive at the end of every lifestage except the last one.

To reiterate, the assumptions on bonds are: there are only one-period bonds, each with a nominal value of \$1, and these pay interest at the *gross* nominal rate r per period. Money is the only medium of exchange, so that commodities can be traded only against money and not bonds, and bonds can be traded only against money and not commodities. There is only one exchange (except for the rollover of the one-period bonds) per period, so that for

5 This is because the fallow periods would be greater for shorter term bonds than for longer term bonds.

net saving in t to be invested in bonds involves the conversion of commodities into money in t, holding money from t to $t+1$ and then converting it into bonds in $t+1$. The reverse process also requires two periods to convert from bonds into money and then from money into commodities. There are no transactions costs (other than the purchase price) of acquiring either bonds or commodities. Endowments are received only in the form of commodities, not of money or bonds.

The symbols are as defined in Chapters 21 and 22. b is the number of bonds, each with a nominal value of \$1. Both m and b are in nominal terms so that their respective real values are m/p and b/p, while c is real consumption and s is real saving. We have assumed in the following that interest at the constant gross rate r on bonds is received in the form of money, which could be used to buy commodities in the period in which it is received.[6] The constraints on the individual, as of the decision period t and with a life of $T+1$ lifestages, are as follows:

For lifestage t:

$$p_t c_t + m_t = p_t w_t \qquad c_t < w_t \tag{1}$$

For lifestage $t+1$:

$$p_{t+1} c_{t+1} + m_{t+1} + b_{t+1} = p_{t+1} w_{t+1} + m_t \tag{2}$$

$$b_{t+1} \leq m_t \tag{3}$$

For lifestage $t+i$, $1 < i < T-1$:

$$p_{t+i} c_{t+i} + m_{t+i} + b_{t+i} = p_{t+i} w_{t+i} + m_{t+i-1} + r b_{t+i-1} \tag{4}$$

$$p_{t+i} c_{t+i} \leq p_{t+i} w_{t+i} + m_{t+i-1} + (r-1) b_{t+i-1} \tag{5}$$

$$b_{t+i} \leq r b_{t+i-1} + m_{t+i-1} \tag{6}$$

Now, moving to the last two lifestages, we have for lifestage $t+T-1$:

$$p_{t+T-1} c_{t+T-1} + m_{t+T-1} = p_{t+T-1} w_{t+T-1} + m_{t+T-2} + r b_{t+T-2} \tag{7}$$

$$p_{t+T-1} c_{t+T-1} \leq p_{t+T-1} w_{t+T-1} + m_{t+T-2} + (r-1) b_{t+T-2} \tag{8}$$

For lifestage T:

$$p_{t+T} c_{t+T} = p_{t+T} w_{t+T} + m_{t+T-1} \tag{9}$$

Demand functions for money and bonds

In the two-lifestages OLG model, the demand for money was positive only in the young lifestage and equaled saving in that lifestage. This was also the amount of the intended dissaving in the following lifestage. In the $T+1$ lifestages model, the demand for real

6 Alternatively, we could have assumed that this interest receipt could only be spent one period later. None of our results depend upon this assumption.

balances is identical with positive saving only in the initial lifestage t (as in equation 1) and with the intended dissaving in the penultimate lifestage $(t+T-1)$ (equation 8). In lifestage $t+i$, $1 < i < t+T-1$, real money demand will equal saving during that lifestage or the intended dissaving in the following lifestage $t+i+1$. There is no demand for money or bonds in the final lifestage T.

For any lifestage $t+i$, the demand for real balances of any given individual will be the greater of (a) saving in the current lifestage in periods with positive saving, (b) expected dissaving in the next lifestage. That is:

$$m^{\mathrm{d}}{}_{t+i}/p_{t+i} = \max(s_{t+i}, |s_{t+i+1}|) \qquad s_{t+i} > 0, s_{t+i+1} < 0 \tag{10}$$

where s stands for saving. The demand for bonds in lifestage $t+i$ equals the individual's real accumulated assets a_{t+i} in lifestage $t+i$ less his demand for real balances. That is,

$$b^{\mathrm{d}}{}_{t+i}/p_{t+i} = a_{t+i} - m^{\mathrm{d}}{}_{t+i}/p_{t+i} \qquad a_{t+i} = \sum_{j=0}^{i} s_{t+j} \tag{11}$$

Leaving aside the "ancillary" constraints – that is, other than the budget constraints – assume for analytical convenience that the desired consumption in any lifestage $t+i$ is always less than the sum of the endowments in $t+i$ and the accumulated savings by $t+i$, $i = 0, 1, \ldots, T-1$. Under this assumption, the individual will always have positive net worth except in the last lifestage $t+T$, so that there would not be any need for the individual to borrow for dissaving, and he would not issue bonds. Consequently, if all individuals in the economy had positive net worth in all lifestages except the last one, there would not exist private bonds issued by households in such an economy. Bonds, like fiat money, would have to be exogenously created by the state – or by firms that could issue bonds to raise capital.

The life cycle consumption hypothesis (LCCH) implies that consumption will be a constant proportion of the present discounted value of lifetime endowments W_t and is specified by:

$$c_{t+i} = k(.)W_t \tag{12}$$

where $k(.)$ is the annuity rate (and kW_t is the income) from W_t, with W_t not being an argument in $k(.)$. Hence, saving s_{t+i} would be:

$$s_{t+i} = w_{t+i} - k(.)W_t \tag{13}$$

From (10) and (13),

$$m^{\mathrm{d}}{}_{t+i}/p_{t+i} = \max(w_{t+i} - k(.)W_t, |w_{t+i+1} - k(.)W_t|) \qquad s_{t+i} > 0, s_{t+i+1} < 0 \tag{14}$$

Aggregate demand for money

The *aggregate* demand for real balances in lifestage $t+i$ is the sum of the real balances demand of two groups: those with positive saving in $t+i$, with this saving being greater than the expected dissaving in $t+i+1$; and those with higher expected dissaving in $t+i+1$. This aggregate demand therefore depends on the time path of endowments over a lifetime, the age

structure of the population, the distribution of each period's endowments in the population, the path of prices over time and the interest rate. Hence, the general form of the nominal aggregate demand function for money, M^{d}, in period $t+i$ would be:

$$M^{\mathrm{d}}{}_{t+i}/p_{t+i} = m^{\mathrm{d}}(p_t, p_{t+1}, \ldots, p_{t+T}, r, \omega_{t+i}, \eta_{t+i}) \tag{15}$$

where ω_{t+i} is the vector of past, present and future endowments of all individuals alive in the economy in period $t+i$, η_{t+i} is the vector of the size of each cohort alive in period $t+i$, and r is the coupon rate on bonds. (15) is homogeneous of degree zero in all prices.

The aggregate demand for bonds in period $t+i$ would be the accumulated wealth of the economy less the aggregate demand for real balances.

Coexistence of money and bonds

Note that money and bonds are not perfect substitutes in the preceding model: money, but not bonds, can be directly traded against commodities; bonds will not be held, but money will have to be held, if consumption is postponed by less than three periods.

In the general case, the aggregate demand for both real balances and bonds would be positive for each period $t+i$, since some members of the population would be saving or expect dissaving in the following period (in which case they hold money), while others would be saving or carrying forward accumulated bonds for consumption several periods into the future (in which case they hold bonds). Both money and bonds would coexist in the economy, since money possesses perfect liquidity as well as acting as a temporary abode of purchasing power, while bonds do not possess liquidity, are a longer term store of value and yield a positive net rate of interest.[7]

Given the above derivation of a positive money demand and a positive finite supply of money to the economy, the price of commodities would be non-zero and the value of money would be positive for each period $t+i$. It is not our intention to derive the specific demand and supply functions, and thereby derive the time path of the value of money for the extended model. The intention was to show, as has been done, the coexistence of money and bonds and a positive value of money for this general case.

Evaluating the implied money demand function

The preceding extended $(T+1)$ period OLG model, like the two-period one, implies a close relationship between saving/dissaving and the demand for money, which is clearly unrealistic. Instead, in a realistic model (see the stylized facts at the beginning of Chapter 21), where money functions as the medium of payments, the model should show a close relationship between the demand for money and consumer expenditures or total expenditures. Further, the close relationship between the demand for money and saving or dissaving in the OLG models is especially suspect for a financially advanced economy where savings can be held

7 If bonds did not pay a positive rate of interest, money would dominate bonds by virtue of its liquidity, so that bonds would not be held. Conversely, if bonds were assigned the same degree of liquidity as money, commodities would have to be directly exchangeable for bonds or indirectly through the intermediation of money but otherwise simultaneously and without transactions costs. In these cases, interest-bearing bonds would dominate money (which does not pay interest), so that money would not be held.

in a variety of marketable and highly liquid "bonds" rather than in currency, demand deposits and savings deposits.

23.1.3 W–M–M theorem in the extended OLG cash-in-advance money–bonds model

The W–M–M theorem was proved in Chapter 22 in the context of a technology in which real balances and the stored commodity or bonds were *perfect substitutes* with *identical* rates of return. However, the extended T + 1-period OLG model with money and bonds set out in the previous subsection assumes a technology at the other extreme: money can be traded against commodities while bonds cannot be directly traded against commodities, so that there is essentially only one trade per period. Hence, bonds cannot substitute for money in the individual's portfolio for the postponement of consumption by less than three periods. Further, for the postponement of consumption by more than three periods, an optimal diversified portfolio of money and bonds will be chosen, with this combination dominating in return over an exclusively money or bond portfolio and over other combinations of money and bonds.[8] This optimal combination will depend upon the technology of exchange and the return on bonds relative to that on money.

To investigate the applicability of the W–M–M theorem in the context of our extended T-period model with money and bonds, we now assume for this model that the commodity cannot be stored, or, if storable, has a rate of return less than on money and bonds, so that no one would use it for storage. Also, in order to allow open market operations between money and bonds, add to this model the assumption that the central bank and the public can trade bonds against money in any period, so that the individual's purchase of bonds with money and the central bank's repurchase of some of these bonds from the individual can be accomplished in the same period.

The distinguishing aspect of the extended model was its cash-in-advance restriction so that the individual could not buy bonds in the first lifestage t of his life. In this model, his demand for bonds in lifestages $T-1$ and T was also zero. His demand for money equaled his saving or dissaving in lifestages t, $T-1$ and T. The individual could hold money or/and bonds in other lifestages, depending upon his saving and dissaving pattern. For the extended model, the central bank's open market transactions with this individual will be subject to these limitations.

To simplify the argument, assume that $t = 0$ and $T = 3$ – that is, with four lifestages from 0 to 3 – and that the net rate of return on bonds is positive while that on money is zero. Further, let there be net saving in lifestages 1, 2 and 3, with a corresponding amount of dissaving in lifestage 4. Bonds are then bought by the individual in lifestage 2 equal to the individual's saving in lifestage 1 and held from lifestage 2 to lifestage 3. Now assume that in lifestage 2 the central bank purchases some of these bonds – in the amount db_2 from the individual against nominal balances of dm_2. To hold government consumption and the individual's lifetime income constant, also assume that the interest on these bonds while in the central bank's possession is returned as a lump-sum transfer to the individual at the maturity of the bonds in lifestage 3. Would the W–M–M theorem hold in this context?

8 Assuming a gross rate of (riskless) return on bonds greater than on money, and for a desired i-period postponement of consumption of one unit from period t to $t+i$, $i > 3$, the unit will be held in bonds from the $(t+1)$th period up to (but not including) the $(t+i-1)$th period, and in money for the t and $(t+i-1)$ periods.

Let the symbols for the demands, supplies and prices prior to the open market operations be designated without an asterisk, but with an asterisk after the operations. Since the operation is conducted in lifestage 2, our explicit analysis will be only for this lifestage. Under the above assumptions, the real demand for money and bonds in lifestage 2 is:[9]

$$b^{\mathrm{d}}{}_2/p_1 = s_1 \tag{16}$$

$$m^{\mathrm{d}}{}_2/p_2 = s_2 \tag{17}$$

Assuming initial equilibrium in the economy with the per capita demands and supplies of the money and bond holders equal to the above demands, we have:

$$b^{\mathrm{s}}{}_2/p_1 = s_1 \tag{18}$$

$$m^{\mathrm{s}}{}_2 = p_2 s_2 \tag{19}$$

From (19) the price level p_2 in period 2 equals $m^{\mathrm{s}}{}_2/s_2$, where $m^{\mathrm{s}}{}_2$ is the exogenously given per capita supply of fiat money. In period 2, the central bank's open market purchase of db_2 of bonds in exchange for dm_2 of money from the individual changes these supplies to:

$$b^{\mathrm{s}*}{}_2 = b^{\mathrm{s}}{}_2 - db_2 \tag{20}$$

$$m^{\mathrm{s}*}{}_2 = m^{\mathrm{s}}{}_2 + dm_2 \tag{21}$$

where $db_2 = dm_2$. These supplies become the individual's holdings of bonds and money. The question is whether the individual will continue to hold them willingly or seek to shed some amount of one asset for the other asset. For the following argument, note that the individual will receive the interest on db_2 as a lump-sum transfer in lifestage 3, but would not associate this as a return on the money increase, so that he does not perceive there to be any interest payment on his original money holdings or on the increase in them.

Since bonds pay interest and money does not, the individual will benefit by immediately reinvesting dm_2 (obtained through the central bank's open market operations) in bonds in the open market. That is, his demand functions for bonds and money in period 2 do not change. Hence:

$$b^{\mathrm{d}*}{}_2/p_1 = s_1 \tag{16$'$}$$

$$m^{\mathrm{d}*}{}_2/p^*{}_2 = s_2 \tag{17$'$}$$

so that, for money market equilibrium in period 2, (21) and (17′) imply that:

$$m^{\mathrm{s}}{}_2 + dm_2 = p^*{}_2 s_2 \tag{22}$$

which implies that:

$$p^*{}_2 = (m^{\mathrm{s}}{}_2 + dm_2)/s_2 \tag{23}$$

9 We have omitted the subscript t in the following, since it is an unnecessary appendage in this section.

so that $p^*_2 > p_2$ for $dm_2 > 0$. That is, since lifestage 2 has an unchanged real-money demand set by saving but an increase in the money supply, the commodity price level will rise and the value of money will fall in lifestage 2. Comparing (16′) and (20), since there is a decrease in the supply of bonds but an unchanged real demand, the price of bonds will rise in lifestage 2.[10] Hence, an expansionary open market operation will increase the price level and the price of bonds. This is contrary to the W–M–M theorem's assertion that the price level and the price of bonds remain unaltered by any open market operation. Therefore, even without bothering to investigate the possible effects in the periods after period 2, we have shown that the W–M–M theorem does not hold in our extended cash-in-advance OLG model.

Another cash-in-advance model

One extension of the preceding model would be introduce labor into it. For this, we now assume that the endowments are only in the form of labor. These are positive in the young/working part of the individual's life and none during his retirement. Each young individual supplies one unit of labor per period and is paid his wage in money in the same period. He can buy commodities for his consumption in the same period or buy bonds if he wishes to carry forward purchasing power to the following periods.

While we leave it to the reader to derive his demand for money and bonds, it is fairly obvious that the individual's demand for money would be more closely related to his consumption expenditures and his demand for bonds would be more closely related to his saving. We also leave it to the reader to show that the strong implications of the W–M–M theorem for open market operations do not hold in this model.

Reappraisal of the W–M–M theorem

The preceding cash-in-advance OLG models allow us to conclude that the W–M–M theorem of open market operations depends upon the specific assumptions introduced for the existence of an alternative asset to money. In this respect, the open market operations against bonds in our extended model seem more realistic in modern economies than against the stored commodity or bonds in the benchmark OLG model without a cash-in-advance constraint. The divergence in the conclusions in the extended model from the W–M–M theorem depends on the former's cash-in-advance constraint, which is not acceptable to many proponents of the OLG framework. Whether such a constraint should or should not be appended to the OLG models is a matter of dispute. But it can be taken as beyond dispute that the OLG models need to be made more realistic. An essential requirement in this revision is that they must incorporate a liquidity differential between fiat money (and deposits in commercial banks) and bonds. This can be achieved through a cash-in-advance constraint, as was done above in the extended model, or through the introduction of the MIUF concept directly or indirectly in the utility functions of the OLG framework – as will be done in the next section. Either one of these modifications of the OLG framework would eliminate the W–M–M theorem from the modified model. Neither seems to be acceptable to many proponents of the OLG framework.

10 Further, since the yield on bonds varies inversely with their price, it will fall in period 2 in response to the decreased supply of bonds brought about by the open market operations.

23.2 An extended OLG model with payments time for purchases and the indirect MIUF

Does money belong in the utility function?

The MIUF models emphasize the medium-of-exchange role of money, with its use as a repository for saving (temporary abode of purchasing power) being an ancillary role in the modern economy. These models and their justification were more elaborately laid out in Chapter 3. For purposes of review, we briefly repeat here some of the justifications for putting money in the utility function. The simplest one is that the individual's utility function includes any good of which more is preferred to less, or vice versa. These preferences represent a subjective decision and take the macroeconomic, social, legal and other aspects of the environment as given. Further, the term "utility" in preference theory does not mean "satisfaction." For a smoker, cigarettes (even though known to him to be harmful to his health and without any objective benefits) are in the utility function. Further, the environment plays an important role in determining the utility function of individuals living in that environment. A social environment in which it was considered normal, sometimes even sophisticated, to smoke led more individuals to have smoking in their utility function than a social environment in which it was frowned upon. The (physical) air pollution environment of a city determines whether clean air appears in the utility function of its residents. The preferences for particular shapes, clothes and designs of clothes usually depend on the "social" or "fashion" environment.[11] Hence, the environment is a major determinant of what the arguments of utility functions are and the utility assigned by individuals to those arguments. Similarly, in a monetary environment in which the very many different commodities and bonds can be exchanged only for money, an individual who wants to trade prefers having more of money balances to less, *ceteris paribus*.

The preceding arguments raise the question: if money is intrinsically useless – that is, in terms of direct usage as a consumer good – as the OLG proponents contend, how does it still yield utility? The above arguments based on the MIUF approach accept this intrinsic uselessness of money, as of many commodities with positive demands and prices in the real-world economies. Nevertheless, they imply attribution of utility (to reiterate, this means only a preference for more rather than less) to money by virtue of the economic environment that makes it more convenient and pleasant to pay for purchases of commodities and bonds with money than with other commodities and bonds. This monetary and economic environment is part of the social set-up, so that money is often called a *social contrivance*. Further, this social set-up is peculiar to each society and economy, so that different economies usually have different goods that act as media of payments in them. For example, the economic environment in Canada makes the Canadian currency, but not the Indian rupee, the medium of payments in Canada and puts it (not the rupee) in the Canadian residents' utility function. The converse argument applies to the use of the rupee in India.

To conclude, the intrinsic usefulness or uselessness of goods in direct consumption is less relevant than the physical, social and economic environment in determining whether or not a

11 For example, ladies' fashionable dresses of the 1890s would have been in the utility function of the ladies in the 1890s, but, being now out of fashion, are not likely to be in the utility function of women in the 1990s. Similar arguments apply to diamonds, body tattoos, powdered wigs, ladies' hats, etc. Are any of them intrinsically useful and must belong in the utility function? Are any of them intrinsically useless and must not be in the utility function?

particular good is in the individual's utility function. The concept of the intrinsic uselessness of money is, therefore, a red herring – and should be discarded.

23.2.1 OLG model extended to incorporate money indirectly in the utility function (MIIUF)

A major rationale given for the use of OLG models is that money does not directly yield consumption services to the individual and should not appear as an argument in the utility function. Consistent with this standpoint, this section starts with a utility function without money. For brevity, it reverts to the two lifestages version of the OLG framework and dispenses with the explicit use of the cash-in-advance constraint. It introduces into this OLG framework the notion introduced in Chapter 3 that the use of money saves the individual's time in carrying out the payment aspects of transactions and that the individual attaches positive utility to leisure. Further, it assumes that there exist one-period bonds with a positive gross coupon rate r, with $r > 1$, while money has the gross rate of 1.

A rather unrealistic assumption of our model in the preceding analysis was that there can be only one transaction among the commodity, money and bonds per period. We will now dispense with this assumption and assume instead that the commodity can be sold for money, which can then be used for buying bonds within the same period. We now assume that there are numerous, heterogeneous commodities in the economy.

Under these assumptions, the individual's two-lifestages intertemporal utility function would be:

$$U(.) = U(c^{y}{}_{t}, h'^{y}{}_{t}, c^{o}{}_{t+1}, h'^{o}{}_{t+1}) \tag{24}$$

The *time constraint* for each lifestage is:

$$h'_{t+i} + n^{\sigma}{}_{t+i} = h_0 \qquad i = 0, 1 \tag{25}$$[12]

where:

h' = leisure
n^{σ} = "transactions time," i.e. time required for making payments for transactions
h_0 = exogenous constraint on the total time (hours) available per period.

The justification for assuming that payments for transactions can require time and reduce leisure is obvious, and was given in Chapter 3. These assumptions imply that consumption and leisure have positive marginal utility but the time required for making payments for transactions has negative utility.[13] In a monetary economy, when the buyer has money to offer for payment, the time spent in making payments is very short and hardly noticeable. However, suppose the buyer has inadequate money balances (or acceptable credit cards)

12 Since endowments are exogenously given to the individual, we have assumed here that none of the time is spent working. This is an unusual assumption. The introduction into the analysis of the labor supplied would not change the results, but would require the introduction of a wage rate for it. This would be an unnecessary encumbrance at this point.

13 This is not the time taken up in shopping, much of which is in selecting the right product to buy. Some years ago, one of my students pointed out that shopping, including window-shopping, was a pleasurable way of spending one's time. Our model does not explicitly include shopping time but can be modified to do so, in addition to leisure and the time devoted to working.

to pay for all his purchases and, instead, tries to offer some commodities, bonds or credit cards unacceptable to the seller in exchange for the remainder of his purchases. The seller may take some time to decide whether to accept these or refuse to accept. In the latter case the buyer has to give up the purchase and try another seller. The time lost in successfully paying for the purchases constitutes n^{σ}. The public's general attitude towards such use of time is that it is irritating and the shorter it is, the better. Our model is consistent with this attitude in attributing negative marginal utility to the time spent in successfully paying for purchases.

The general form of the function for transactions time (i.e. paying for purchases) assumed in Chapter 3 was:

$$n^{\sigma}{}_{t+i} = n^{\sigma}(m_{t+i}/p_{t+i}, c_{t+i}) \tag{26}$$

where the partial derivative with respect to the second argument (i.e. c_{t+i}) is positive and the first one (m_{t+i}/p_{t+i}) is negative: for a given amount of purchases, holding more money balances up to the total cost of the purchases reduces the time spent in making payments for them. From (24) to (26),

$$U(.) = U(c^{y}{}_{t}, h_0 - n^{\sigma}(m^{y}{}_{t}/p_t, c^{y}{}_{t}), c^{o}{}_{t+1}, h_0 - n^{\sigma}(m^{o}{}_{t+1}/p_t + 1, c^{o}{}_{t+1}) \tag{27}$$

where (27) is the indirect intertemporal utility function. Real balances appear in the utility function for both lifestages. Since the individual dies at the end of the second lifestage and any money balances still held by him would have no value for him, we assume that he is able to rent from others the (usage of) money balances in the second lifestage.[14] Since the alternative asset to money is the illiquid asset, bonds, with a gross coupon rate r, this rental rate of money in perfectly competitive and efficient markets, without transactions costs, would be $(r - 1)$ per period. He could also rent the money balances during the first lifestage.[15] However, we assume that he benefits by owning them outright in this lifestage since he needs to carry over purchasing power through holding them to the second lifestage. Hence, money balances now perform both transactions and savings roles in all periods except the last one, when they act only as a medium of payments. In this final period of life, the individual will not own money or bonds since their amount left over at the end of the period would be of no use for him.

Therefore, the assumptions are that the saving in the first lifestage is greater than or equal to the demand for money therein and that the money balances in the first lifestage are owned outright but are rented in the second/terminal lifestage. We also assume, as we did earlier for the benchmark OLG model, that the first lifestage endowment exceeds its optimal consumption.

14 This assumption is not needed if the individual's lifetime is infinitely long.

15 This assumption, that the services of money are only rented, is necessary only for the terminal period. In a T-lifestage model, this assumption would be needed only for period T.

The alternatives of owning versus renting money are similar to those of owning versus renting a consumer durable good such as a refrigerator, with the user cost from owning equaling the rental rate in perfect markets. Chapter 7 has discussed this concept in greater detail.

Since both money and bonds can now be bought in period 1, the budget constraint for the first lifestage becomes:

$$pc^{y}{}_{t} + m^{y}{}_{t} + b^{y}{}_{t} = p_t w^{y}{}_{t} \tag{28}$$

The second lifestage budget constraint is:

$$p_{t+1} c^{o}{}_{t+1} + (r-1) m^{o}{}_{t+1} = p_{t+1} w^{o}{}_{t+1} + m^{y}{}_{t} + r b^{y}{}_{t} \tag{29}$$

where money is held in both lifestages but bonds are held only in the first lifestage. Further, money is owned outright in the first lifestage at a cost of unity but is rented in the second lifestage at a rental cost of $(r-1)$. From (29),

$$r b^{y}{}_{t} = p_{t+1} c^{o}{}_{t+1} + (r-1) m^{o}{}_{t+1} - p_{t+1} w^{o}{}_{t+1} - m^{y}{}_{t} \tag{30}$$

Substituting (30) in (28) to eliminate $b^{y}{}_{t}$ gives the individual's lifetime budget constraint as:

$$p_t c^{y}{}_{t} + \{p_{t+1}/r\} c^{o}{}_{t+1} + (1 - 1/r) m^{y}{}_{t} + \{(r-1)/r\} m^{o}{}_{t+1} = p_t w^{y}{}_{t} + \{p_{t+1}/r\} w^{o}{}_{t+1} \tag{31}$$

The individual's real lifetime wealth (i.e. the present discounted values of lifetime endowments) W_t is now given by:

$$W_t = w^{y}{}_{t} + (p_{t+1}/p_t)(1/r) w^{o}{}_{t+1} \tag{32}$$

Intertemporal utility maximization – with respect to c_t, c_{t+1}, m_t and m_{t+1} – of (27) subject to (31) yields the per capita demand functions for real balances of the form:

$$m_t/p_t = \phi_t(p_t/p_{t+1}, r, W_t) \qquad i = 0, 1 \tag{33}$$

The per capita real demand for bonds would be:

$$\begin{aligned} b^{y}{}_{t}/p_t &= w^{y}{}_{t} - c^{y}{}_{t} - m^{y}{}_{t}/p_t = w^{y}{}_{t} - c^{y}{}_{t} - \phi_t(p_t/p_{t+1}, r, W_t) \\ &= \Theta_t(p_t/p_{t+1}, r, W_t, w^{y}{}_{t}) \\ b^{o}{}_{t+1} &= 0 \end{aligned} \tag{34}$$

Note that the demand for bonds, but not for money, in the second lifestage is zero in the current two-lifestages model.

Saving in the two lifestages would be:

$$\begin{aligned} p_t s^{y}{}_{t} &= p_t w^{y}{}_{t} - p_t c^{y}{}_{t} = m_t + b_t > 0 \\ s^{o}{}_{t+1} &= 0 \end{aligned} \tag{35}$$

Hence, in the general case, there would be positive demands for real balances in both lifestages, a positive demand for bonds in the first one only, and the demand for real balances would differ from that for saving. In particular, with $b^{y}{}_{t} > 0$, the demand for real balances would be less than saving in lifestage 1, and – with $m^{o}{}_{t+1} > 0$ but $s^{y}{}_{t} = 0$ – greater than

saving in lifestage 2. This model is, therefore, somewhat more realistic than the basic OLG model in which money was merely a medium for saving, the demand for real balances had to be identical to saving, and in which money and bonds under the above assumptions could not simultaneously have a positive demand.

We do not carry this analysis further to investigate the value of money, the efficiency of monetary stability or that of monetary growth, etc., as has been done for the benchmark OLG model. However, it should be clear from a comparison of the above analysis with that of the benchmark OLG model that the money demand functions can be dominated,[16] as in period 2, by the transactions role of money rather than by its medium-of-saving role, in which money is usually dominated by coupon-bearing bonds. Further, the demand for money is positive even when the individual's saving is zero, as in the second lifestage.

In the economy-wide context, the positive demand for money in each period for transactions implies that while the current value of money does depend on future prices, it would remain positive even if the value of money at some future date was expected to be zero. Hence, money would continue to be used even in extreme hyperinflations in monetary economies, since money continues to be required for convenience and reduction in the transactions time when purchasing commodities.

An illustration

The above analysis can be illustrated by a time-additive log-linear form of the intertemporal utility function. That is, let:

$$U(.) = u^{y}{}_{t} + \delta u^{o}{}_{t+1} \tag{36}$$

Specify the period utility function as:

$$\begin{aligned} u_{t+i} &= \ln c^{\rho}{}_{t+i} h'^{\gamma}{}_{t+i} \qquad i = 0, 1 \\ &= \ln c^{\rho}{}_{t+i} \cdot (h_0 - n^{\sigma}(m_{t+i}/p_{t+i}, c_{t+i}))^{\gamma} \end{aligned} \tag{37}$$

where:

$U(.)$ = intertemporal utility function
$u(.)$ = period utility function
$\delta = 1/r$
r = gross rate of time preference (gross subjective discount factor).

Assume that the endowments are of time h_0 and that the technology of transactions is a Cobb–Douglas one, such that the time h'_t available for leisure (i.e. excluding transactions time)[17] is:

$$h'_t = h_0 - n^{\sigma}(m^{y}{}_{t}/p_t, c_t) = k(m_t/p_t)^{\alpha} c^{-\beta}{}_{t} \qquad k \geq 0 \tag{38}$$

Hence,

$$u_{t+i} = \ln[kc^{(\rho-\gamma\beta)}{}_{t+i} \cdot (m_{t+i}/p_{t+i})^{\gamma\alpha}] \qquad i = 0, 1 \tag{39}$$

16 This dominance is likely to increase as the number of commodities increases and as the minimum period for which bonds have to be held decreases.

17 Note that in this economy, with endowments of commodities but no production or employment, leisure equals total available time less transactions time.

From (36) and (39),

$$U(.) = \ln[kc^{(\rho-\gamma\beta)}{}_t \cdot (m_t/p_t)^{\gamma\alpha}] + \delta \ln[kc^{(\rho-\gamma\beta)}{}_{t+1} \cdot (m_{t+1}/p_{t+1})^{\gamma\alpha}]$$

which gives:

$$\begin{aligned} U(.) = {} & [\ln k + (\rho - \gamma\beta)\ln c_t + \gamma\alpha \ln(m_t/p_t)] \\ & + \delta[\ln k + (\rho - \gamma\beta)\ln c_{t+1} + \gamma\alpha \ln(m_{t+1}/p_{t+1})] \end{aligned} \tag{40}$$

The individual is assumed to maximize (40), subject to the budget constraint (31), with respect to c_t, c_{t+1}, $m^r{}_t$ $(= m_t/p_t)$ and $m^r{}_{t+1}(= m_{t+1}/p_{t+1})$, where m^r signifies real balances held by the individual. Note that bonds are not a variable in this system, though they were in the period budget constraints (28) and (29), so that the optimal demands for bonds in each lifestage will be derived from these constraints for the derived optimal values of consumption and real balances.

The Lagrangian function for (40) subject to (31) is:

$$\begin{aligned} L = {} & [\ln k + (\rho - \gamma\beta)\ln c_t + \gamma\alpha \ln(m^r{}_t)] + \delta[\ln k + (\rho - \gamma\beta)\ln c_{t+1} + \gamma\alpha \ln(m^r{}_{t+1})] \\ & - \lambda[c^y{}_t + \{p_{t+1}/p_t r\}c^o{}_{t+1} + (1 - 1/r)m^r{}_t + \{(r-1)/r\}m^r{}_{t+1} \\ & - w^y{}_t - \{p_{t+1}/p_t r\}w^o{}_{t+1}] \end{aligned}$$

where λ is the Lagrangian multiplier. Maximization of this function with respect to c_t, c_{t+1}, $m^r{}_t$, $m^r{}_{t+1}$ and λ will yield the Euler conditions, which can be solved for the optimal values of these variables. We leave it to the interested reader to do so and show that these optimal values imply that:

$$m^r{}_t < s_t$$

$$m^r{}_{t+1} > s_{t+1}$$

Further, derive p_{t+i} and v_{t+i}, for $i = 1, 2$. Also derive the demand functions for bonds in each period.[18]

This model and its implied demand functions are closer to the ones that underlie the IS–LM macroeconomic model's demand functions for commodities, money and bonds. In a still more general context, the budget and time constraints can be modified to allow usage of time for work, leisure and payments for transaction. Depending upon the forms of the assumed functions, this extension would imply that the use of money by households increases

18 There should be a positive demand for real balances in each lifestage, including the terminal one, and a positive demand for bonds only in the first lifestage but a zero demand for bonds in the (terminal) second lifestage. Since there can be a positive demand for money for transactions purposes in each period, its value in each period need not be zero even if its value in a future period is expected to be zero. This is clearly a much more realistic result than the demand for money generated by the benchmark OLG model.

their labor supply and, therefore, the output produced in the economy by the increase in employment.[19] But they still would not take account of the impact of the firm's holdings of money on the firm's production of commodities. This is attempted in the next section.

23.3 An extended OLG model for firms with money indirectly in the production function (MIIPF)

23.3.1 Rationale for putting real balances in the production function

The preceding chapter claimed that it was a major defect of the benchmark OLG models that, while the introduction of money in them improved allocative efficiency in the consumption of a given path of endowments, it did not make for an increase in the output and availability of goods for consumption along that path. This runs against the dominant stylized facts of monetary economics that a monetary economy produces a greater output than a barter or autarchic one. One way of obtaining such a result from the OLG models is to introduce a production sector. This requires introducing profit-maximizing firms, with at least labor inputs in production.

For the simplest, two-lifestages OLG model with production, assume that each individual is endowed not with commodities but with labor time. He sells it to firms that produce commodities with it. Neither labor nor commodities can be stored. Firms use labor to produce commodities; there is no other input. In particular, the firm's real balances are not an input in this production function.

Under a barter arrangement, the workers have to be paid their (real) wages in the form of the output of the commodities they helped to produce. If the worker's wage in the first lifestage exceeds his optimal consumption, then, under the assumptions of the benchmark OLG models with money, any saving will be carried over to the second lifestage in the form of money. Since this barter scenario does not allow any benefits to the representative firm from the usage of money by it, it would not hold money nor would such balances increase its output. This is counterfactual and we need to modify the OLG model further for realism. The model is also unrealistic in its assumption that firms pay their workers in the commodity they help to produce, without such a practice reducing the supply of labor to it, and therefore its own employment and output.

Since the proponents of OLG models are generally against putting money directly in the consumer's utility function, they are also against inserting money directly into the firm's production function. The argument given for this is that fiat money is intrinsically useless and does not directly contribute as an input in production. Consistent with this standpoint, we have assumed above that money does not directly enter the production function. However, even if we accept this, in a monetary economy, while all input suppliers are willing to accept money in exchange for their inputs, they are not all willing to do so against the commodities produced with their inputs. Further, in the monetary economy, while all consumers are willing to pay for their purchases in money, any given consumer would have difficulty in finding precisely that supplier who would want the input that the consumer can supply and in exactly the right amount.

19 This is an important effect of the use of money. We leave it to the reader to show that adequate money holdings increase labor supply to firms for production relative to inadequate amounts, and thereby increase the economy's output. Chapter 3 provides the relevant analysis.

An island parable

An "island parable" with heterogeneous commodities, firms and consumers to illustrate the above scenario is as follows. The economy has n islands. There are n commodities and n firms, with each firm producing a particular commodity and located on its own separate island. Each island has consumers/workers, with the workers on an island demanding a particular commodity for consumption. If the firm is neither willing to pay wages in money nor willing to accept payments for sales in money, it must send out employees, "emissaries," to search for islands with workers who would accept wages in its commodity – conversely, to look for islands with consumers who will pay for the purchase of its commodity with their labor. Let the possibilities of reaching the appropriate islands for effective exchange be randomly distributed. With n commodities and n islands, an emissary has a probability $1/n$ of hitting the right island in one day or other period. If the firm sends out only one emissary per day, then, on the $(n-1)/n$ of the days that he is not successful the firm does not get any inputs and does not produce any output. Further, the worker uncontacted, but nevertheless appropriate, on the appropriate island will not get to sell his labor services and does not receive any wage payment, which decreases his intertemporal consumption of commodities.

Now assume that the firm without money holdings tries to reduce its risks of shutting down or spoiling its commodity by employing a larger number of emissaries. Two emissaries sent out in search of the appropriate workers reduce the possibility of shutdown for lack of workers to $(n-2)/n$. Three such emissaries reduce it to $(n-3)/n$, and so on. As against this benefit, there is the problem that multiple emissaries working independently could end up hiring more workers than can be utilized and the excess workers have to be paid even if they are not put to work, thereby involving losses for the firm.

As against the above scenario, consider an alternative one in which all economic agents use money and stand willing to accept it in exchange for labor and commodities. An emissary in search of workers would be able to hire the required workers from the first island he reaches, promising to pay wages in money. However, the firm would still to need to send out salespersons to find consumers who want to buy its product. With all consumers willing to pay in money, the salesperson with the commodity to sell would be able to sell them on the first island he reaches, with the consumers willing to buy its particular product. The money brought back by the salespersons would be paid to the workers as wages at the end of the period. The workers would use their wages, paid in money, for purchasing their desired commodities in the next period and for saving. Depending on its "effective demand," which is a function of how many salespersons are employed and how many consumers can pay in money, the firm would have its required workers in each period and would not have to shut down for some of the periods. Therefore, employment and output would be higher for the firm using money in addition to emissaries and salespersons, and, in the aggregate, for all firms in the monetary economy when compared with the expected value of employment and output in the non-monetary economy, or in one in which firms and consumers hold inadequate money balances. Hence, the production function for such an economy would include money in the firms' production functions.

Empirical evidence on money in the production function

Our earlier arguments have provided theoretical and intuitive reasons why firms' output should vary with their money holdings. For illustrative purposes, we cite just one empirical

study that supports this contention. While some studies report a finding against the inclusion of money in the production function, Sephton (1988) reports, using annual data from 1938 to 1978 for the United States, that firms' real balances are a valid input in the context of a CES (constant elasticity of substitution) production function (see Chapter 7 for the specific form of this function).

23.3.2 Profit maximization and the demand for money by the firm

Clearly, the profit-maximizing firm in a monetary economy has a tradeoff between the amount of money it uses and the number of emissaries it has to keep. If it keeps more money balances, it needs to use fewer of its employees as emissaries, so that the released labor can be used to produce greater output. To incorporate this tradeoff into the analysis of the representative firm, assume a production technology with labor as the only input, of which n' of the representative worker's time is employed by the firm in production and n'' is used to carry out exchanges (as an emissary in search of other workers who would accept the firm's commodity in exchange for their labor and to sell the firm's output). The firm's total employment of the worker's time in producing and selling output x is $n' + n''$. Assume that the firm's production function *per worker* is given by:

$$x = f(n') \qquad \partial x/\partial n' > 0 \tag{41}$$

Further, assume that the payments technology for the monetary economy in which the firm operates is specified by:

$$n'' = \phi(m^{\mathrm{f}}/p, x) \tag{42}$$

where $\partial n''/\partial(m^{\mathrm{f}}/p) < 0$ and $\partial n''/\partial x > 0$, ϕ is the payments technology function and

x = firm's output
n' = time per worker used directly in production
n'' = time per worker used as an emissary
m^{f}/p = real balances per worker used by the firm
p = price of the firm's product.

Let the firm's total employment of the worker's time be $\underline{n}$. From (41) and (42),

$$x = f(\underline{n} - \phi(m^{\mathrm{f}}/p, x)) \tag{43}$$

where $\phi x/\partial(m^{\mathrm{f}}/p) = -(\partial f/\partial n')\{\partial\phi/\partial(m^{\mathrm{f}}/p)\} \geq 0$, so that the firm's output is higher if it uses more real balances but only up to a limit. (43) is an indirect production function, with the real balances brought indirectly into it, and can be rewritten as:

$$x = x(\underline{n}, m^{\mathrm{f}}/p) \qquad \partial x/\partial(m^{\mathrm{f}}/p) \geq 0$$

where $x(\underline{n}, m^{\mathrm{f}}/p)$ is the firm's indirect production function.

The firm maximizes its profit π, where π for given employment $\underline{n}$ is:

$$\pi = pf(\underline{n} - \phi(.)) - W\underline{n} - \rho_m m \tag{44}$$

where ρ_m is the user cost of nominal balances and W is the nominal wage rate. (44) implies that the firm's supply function for output and its demand function for real balances (m/p) are given by:

$$x^{s} = x^{s}(W/p, \rho_m; \underline{n}) \tag{45}$$

$$m^{fd}/p = \Psi(W/p, \rho_m; \underline{n}) \tag{46}$$

Note that the payments system is not static. An improvement in it makes the usage of real balances more efficient in carrying out transactions and therefore more profitable relative to the use of emissaries, so that it reduces n'', thereby increasing n' and x. Hence, more advanced payments systems increase the economy's output for a given employment level.

If the firm's output also depends, besides its employment of workers directly used for production, only on that part of its capital stock that is directly used in production (but not on the part that has to be used to support its emissaries, e.g. as "rowboats to go among islands"), the use of real balances will allow a reduction in the part of capital that is not used directly in production, leaving more capital – out of a given total amount – for direct usage in production, thereby contributing to an increase in the firm's output.

Section 23.2 illustrated the incorporation of the indirect MIUF approach into the two-lifestages OLG model by assuming Cobb–Douglas functions for the transactions technology. Such functions can also be used for illustrating the indirect MIPF and would assume a unit elasticity of substitution between labor, capital and real balances. Given such a production function, the maximization by the firm of profits would imply positive (up to a point) but declining marginal productivity of real balances, so that the firm would have a positive demand for real balances. In the aggregate, an economy with such firms would produce higher output by having positive real balances, facilitating the hiring of inputs and the sale of the output that is produced in the economy. We leave it to the interested reader to pursue the mathematics of this illustration.

23.3.3 Intuitive empirical evidence

Most economies have had well-established monetary exchanges for very long periods, so that it becomes easy to overlook the benefits from the intermediation of money in making payments. We offer empirical evidence in the form of descriptions of real-world cases where money was not available for making such payments, or only gradually came into existence. One of these descriptions is provided by Radford (1945), who describes the economy of a prisoner of war camp during the Second World War. This was basically an exchange economy, with endowments in the form of prison rations and Red Cross packages and exchanges of a variety of consumer goods such as cigarettes, tinned milk, biscuits, clothing etc. The use of cigarettes as money and eventually paper money (issued by the restaurant and shop in the camp) evolved because they facilitated exchanges. There was thus both a transactions usage of money for obtaining a *better consumption pattern within the same day* than the exogenous endowment system, and a usage of money to change the intertemporal consumption pattern over days through holding savings in the form of money. OLG models of money capture the second usage of money but not the first one. In fact, Radford's article shows that the second usage would never have come into being if the first usage had not arisen, and that most trades were for the first type of usage.

Prisoners were often transferred between camps, sometimes at short notice. Further, there was the general expectation in the camp that the war would end sooner or later and that all prisoners would then be released. Hence, this economy was fully expected to have a *finite life*. At the end of this life, the prison's paper money would become useless and have zero value. The OLG models predict that, under these circumstances, paper money would have zero value in all periods and would never have come into existence. Evidence of the evolution, value and use of paper money in the camp contradicts this implication of the OLG models.

Our second bit of descriptive evidence is given in Exhibit A. This provides a description of a context in which money becomes unavailable for transactions, though the production technology and the physical capital are constant. The exhibit shows that when a firm cannot pay its workers in money, a great deal of the workers' (including their family members') time has to be diverted from labor supply and leisure to transactions activities. Consequently, the labor engaged in production – both market and home – is reduced and so is the economy's output. In addition, there is a very significant loss of utility from the reduction in leisure and in other ways (through being "humiliated").

In addition, the firm had trouble getting its raw materials because the suppliers wanted "real money" – in the terminology used by one of the factory's employees – and not the firm's output for barter. Without money to offer in exchange for raw materials, the factory had to temporarily shut down, thereby reducing its output.

Hence, the conclusion from exhibit A is that the use of money by firms and workers increases their production and utility. It also confirms our contention that it is the monetary and economic environment – in this context, the need to trade with others when there are numerous heterogeneous commodities – that creates the productivity of money and its utility. Their denial by the notion of the "intrinsic uselessness of money" belongs legitimately to the barter economy. For monetary economies, the assertion by the OLG models that money is useless is erroneous and contrary to the fact of its intrinsic usefulness in the monetary environment.

EXHIBIT A

An experiment in reality[20]

Real-world experiments are rarely available to monetary economists for checking on their theories. The switch from the centralized system of the Soviet Union to the capitalist one of Russia provided many instances of the switch from a monetary system to a barter one. In these switches, the production technology and the physical system remained constant, at least for some time, but the mode of payment of workers and other inputs changed from payment in money to payment in kind. What were the consequences of this switch?

One of the instances of such a switch occurred in the context of a glass and crystal factory in Gus-Khurstalny in Russia in 1997. The factory used red lead as an input and had earlier obtained it from sources that were formerly in the Soviet Union but were now not in Russia. These suppliers wanted payment in money and refused to accept payment in the factory's products, or through some other multilateral barter arrangement. Since the factory did not have money to pay, it did not get adequate supplies of red lead and had to periodically shut down.

20 The following is based on an article that appeared in the *Montreal Gazette* (page D2), February 7, 1997.

The workers were, however, willing to take their pay in glass and crystal – thereby agreeing to barter labor services for the factory's products – since other jobs were very scarce in the area. This meant that the workers had to bring adequate transport to the factory and had to make arrangements to sell their "salary-in-kind" or barter it for the products that they needed.

Family members were often drafted to set up stalls by the roadside to try to "market" the salary-in-kind. It was sometimes possible to barter some for fruits and vegetables. Some might be bartered for the products, such as linen and bathrobes, of other factories in the area. There could also be occasional sales to travelers with money who might stop to buy from the workers' roadside stalls. But, besides the considerable labor time involved in these transactions, there were also many unpleasant elements. These included "protection payments" to thugs and the dust, discomfort, taunts and humiliation of being on such stalls. Often, the day's salary-in-kind of a worker could not be sold by a day's work on a stall, so that exchange could take more time than production.

With a shortage of money, the factory also had difficulty in making its payments in cash to the town, so that some form of barter arrangements had to be made. Under these arrangements, the town sent unemployed workers and others directly to the factory for the receipt of their unemployment benefits, paid in glass and crystal, which the recipients then had to sell or trade for the commodities that they needed.

23.4 Basic OLG model with MIIUF and MIIPF

To combine a technology based on indirect MIUF and MIPF with the OLG model, note that each worker has been allocated an exogenous endowment h_0 of time in each period. Focusing on the net amount not taken up in leisure h', some of this time is used by the worker looking for work and in transactions n^σ, and the rest is offered to firms as labor supply n. But if the worker is paid for his labor by the firm in the firm's commodity, rather than in money, he has to spend part of his time (e.g. on the "roadside stand" in Exhibit A) in an attempt to exchange this commodity for the ones he needs for consumption, thereby reducing the labor time he sells to the firm. In fact, his productivity and real wage per hour is reduced by the time devoted on the roadside stand, which is a consequence of his employment in a barter economy. Although this is a very significant element in the illustration in Exhibit A, the following analysis ignores it in order to simplify the argument.

Assuming full employment, n' out of n is used by the firm for emissaries and only the remainder n'' is available for production, so that $n = n' + n''$ and $h_0 = h' + n^\sigma + n' + n''$, where, as a reminder, h_0 is the total available time per worker, h' is leisure, n^σ is the time spent in paying for transactions, n'' is used by the firm for facilitating transactions through emissaries and only n' is directly used in production. Introducing time subscripts for the two-period OLG model, n^σ_{t+i} depends on $(m^h/p)_{t+i}$, while n''_{t+i} depends on $(m^f/p)_{t+i}$, where m^h/p are the workers' own real transactions balances and m^f/p are the firm's real transactions balances.

Let the output produced by n'_{t+i} of time be w_{t+i} and assume that it is fully paid out in wages. Hence, the worker's income/output in the two lifestages will be:

$$\begin{aligned} w^{\mathrm{y}}{}_t &= \Psi(n'^{\mathrm{y}}{}_t) = \Psi(h_0 - h'^{\mathrm{y}}{}_t - n^{\sigma\,\mathrm{y}}{}_t - n''^{\mathrm{y}}{}_t) \\ &= \Psi(h_0 - h'^{\mathrm{y}}{}_t, m^{\mathrm{hy}}{}_t/p_t, m^{\mathrm{fy}}{}_t/p_t) \end{aligned} \tag{47}$$

$$w^{o}{}_{t+1} = \Psi(n'^{y}{}_{t+1}) = \Psi(h_0 - h'^{y}{}_{t+1} - n^{\sigma y}{}_{t+1} - n''^{y}{}_{t+1})$$
$$= \Psi(h_0 - h'^{o}{}_{t+1}, m^{ho}{}_{t+1}/p_{t+1}, m^{fo}{}_{t+1}/p_{t+1}) \tag{48}$$

where $\partial w_{t+i}/\partial(m^{h}{}_{t+i}/p_{t+i}) > 0$ and $\partial w_{t+i}/\partial(m^{f}{}_{t+i}/p_{t+i}) > 0$ for $i = 0, 1$. Consequently, an OLG economy with the indirect MIUF and MIPF produces more output in each period because (a) the use of money by workers increases the labor supply by reducing the time spent by workers in making payments for transactions, thereby increasing employment, and (b) the use of money by firms increases the amount of employed labor devoted to direct production.

In equilibrium, there will be an optimal amount of real balances held by both workers and firms, with the economy's per capita demand for nominal balances being given by:

$$m^{d}{}_{t+i} = m^{h}{}_{t+i} + m^{f}{}_{t+i} + f(p_{t+i}s_{t+i}) \tag{49}$$

where $f(.)$ is the medium of saving – i.e. store of value – demand for real balances in the OLG framework. (49) assumes, for simplification, that $f(.)$ is additive to the other two components of money demand. This total demand for real balances, with some held by workers and some by firms, will not, except as a coincidence in special circumstances, be identical with saving in each period, thereby eliminating one of the undesirable implications of the basic OLG model. This was the close – usually identical – association between saving and the demand for real balances. In fact, according to (49), the transactions demand for real balances by both firms and households will be related to consumption and output, rather than saving, though there is also a component that is related to saving.

While a more advanced payments system may decrease or increase the demand for real balances, it will increase the time available for direct production by economizing on the use of workers' time in transactions and as emissaries, thereby increasing the output produced each period in the economy and the income of the workers in each lifestage. This will increase both the representative worker's welfare and social welfare generally.

As a corollary, saving and the demand for real balances by individuals and firms would depend upon the payment use of money, its cost and the stage of development of the payments system, as well as on the portfolio demand for money (i.e. the desirability of holding savings in money). Further, with money performing a medium-of-payments role during each period – in addition to its store-of-value role across periods – and partly held for that purpose, while bonds are not, an open market operation between money and bonds will reject the W–M–M theorem for fiat money in OLG models. Similarly, with money performing this-medium-of payments role while stored commodities cannot, an open market operation between money and the stored commodity will also not bear out this theorem. In particular, the price level will not be invariant with respect to such open market operations.

Conclusions

This chapter has amended the OLG framework by bringing into it a medium-of-payments role through a cash-in-advance constraint or by putting money indirectly into the utility and production functions. In the modified cash-in-advance model, fiat money can coexist with bonds paying a positive rate of interest. Money dominates bonds as a temporary abode (from one period to the next one) of purchasing power by virtue of its liquidity, while bonds dominate money as a longer-term store of value. However, the individual's demand for

money in any period i is identical with the larger of the saving in period i or the dissaving in period $i+1$, in terms of their absolute values. This implication is unrealistic for the modern financially developed economy where there is rarely much of a correspondence between the quite different variables of real balances and saving.

The second extension of the OLG benchmark model brought in the payments/transactions time hypothesis in the context of a multiplicity of heterogeneous commodities, and implied that money appears in the indirect utility function. Money could now be used not only as a possible vehicle for saving but also for purchasing commodities within each period, while bonds were illiquid, i.e. they could not be exchanged for commodities or other bonds directly, but had a higher rate of return than money. The individual's demand for money in this context became distinct from his saving and dissaving, and depended on his liquidity needs.

The addition of the indirect MIUF hypothesis to the benchmark OLG framework does not change the level of output per employed worker produced in the economy and therefore does not explain why monetary economies have higher output per worker than non-monetary ones. To explain this, we need production in the model. The use of money releases labor from work associated with paying for purchases. The labor thus released becomes utilized in production and increases the productive capacity of the economy as a whole. In addition, the use of money increases the size of the market and specialization in production, so that its use leads to economies of scale, resulting in greater average productivity of labor.

While the "merger" of the indirect MIUF and MIPF hypotheses with the OLG hypothesis yields much more plausible implications, as can be seen by their comparison with the stylized facts, such a "merger" – or is it really a "takeover"? – is unlikely to be acceptable to many proponents of the benchmark OLG models with money. Further, the implications of this merger are very similar to those of the MIUF and MIPF models, so that the distinctiveness of the OLG framework becomes severely diluted. This merger, however, is likely to be quite acceptable to those in the MIUF and MIPF traditions, since they see the OLG framework as one way of extending the MIUF and MIPF notions to an intertemporal context.

Summary of critical conclusions

- Cash-in-advance models of transactions impose a distinction between the liquidity of money and bonds, with money being the more liquid asset, and imply a distinct demand function for money and bonds.
- The use of money by consumers allows a reduction in the transactions time devoted to the sale of their labor and their purchases of commodities, and implies an increase in their labor supply to firms.
- The use of money by firms reduces the labor and capital inputs that have to be allocated to the sale of their output and the hiring of their inputs, and therefore increases the proportion of their employees employed directly in production.
- At the economy's level, the use of money by consumers and firms increases the leisure time and labor supply of workers and the output produced in the economy. Therefore, the use of money in the economy increases both consumer welfare and profits of firms.
- The empirical stylized facts are better explained by OLG models incorporating MIUF (or MIIUF) and MIPF (or MIIPF) than by models without such functions.

Review and discussion questions

1. Which of the following approaches to money demand do you prefer: (a) MIUF, (b) indirect MIUF, (c) cash-in-advance, or (d) pure OLG models of money? Why?
2. Some economists have claimed that the OLG model of fiat money includes what is essential for a good theory of money and that it should be recognized as the best available one. Draw up a list of (a) what you consider to be essential for a good theory of money, and (b) the policy implications of the OLG model. Assess how far (b) satisfies (a).
3. Given the stylized facts on money in Chapter 21, how far does the MIUF approach (i) without time and overlapping generations, (ii) with overlapping generations, satisfy the items in the list (a) that you specified in answer to the preceding question? In answering (i), you may, if you want, use the analysis of Chapter 3, or the IS–LM or IS–IRT framework, in embodying the MIUF approach.
4. What is the meaning of utility in microeconomic theory? In answering this question, consider the following. Why are goods that seem to harm the individual's physical or mental wellbeing placed in the utility function? Why do goods with upward-sloping demand functions (i.e. with $\partial x^{d}{}_{i}/\partial p_i > 0$) due to snob appeal possess the marginal utility that they do? What is the significance of the environment (physical, economic, social, political, etc.) as an element in the determination of the preferences of the individual?
5. Why does the W–M–M theorem hold in the analysis of Chapter 22 but not in the analysis presented in this chapter? What is the critical difference in the assumptions that changed this finding and what is your own assessment of their plausibility?

   Does the W–M–M theorem hold or not hold in modern developed economies? Does it do so in the LDCs with poorly developed financial markets? Discuss.
6. If currency (a fiat money) is intrinsically useless for firms and households, are demand/checking deposits also intrinsically useless for them? Why, then, do they hold these when there are other highly liquid assets such as savings deposits and money market mutual funds (MMMF) that offer higher returns than demand deposits? If the only demand for checking deposits is for holding savings, as in the OLG models of money, why do banks offer such checking accounts and not confine themselves to savings and term deposit accounts? Why do they offer the latter rather than confining themselves to MMMF?
7. Discuss the different parts of the following statement: "For the modern economy, the theory of the speculative demand for money has come to the conclusion that there is no longer an asset demand for M1 and that any demand for it is only for facilitating current transactions. However, the OLG models have chosen to assert the asset demand for fiat money, and presumably for M1 also, to the exclusion of the transactions demand and the inventory models of the transactions demand. The OLG models of money, therefore, incorporate an out-of-date and currently non-existent motive for holding M1."
8. Set up an OLG model with two lifestages, a utility function without money in it and a Cobb–Douglas transactions technology for making payments to complete purchases of commodities. Make any other assumptions that you require. For this model, derive the representative worker's supply function for labor. Discuss the relationship between the worker's supply of labor and the real money balances held by him.
9. Set up an OLG model with two lifestages, a Cobb–Douglas production function with capital and labor but without money, and a Cobb–Douglas transactions technology that uses both capital and labor for making payments to complete purchases of inputs and sales of output. Make any other assumptions that you require. For this model, derive the

representative firm's supply function for output. Discuss the relationship between the firm's output and the real money balances held by it.

10. Set up an OLG model with two lifestages, an indirect MIUF approach with a Cobb–Douglas transactions technology and an indirect MIPF with a unit elasticity of substitution between labor, capital and real balances. Make any other assumptions that you require. For this model, derive the representative firm's demand function for real balances. Does this demand in each lifestage equal saving, or even depend on saving, in that lifestage?

References

Brock, W.A. "Overlapping generations models with money and transactions costs." In B.M. Friedman and F.H. Hahn, eds, *Handbook of Monetary Economics*, vol. I. Amsterdam: North-Holland, 1990.

McCallum, B.T. "The role of overlapping generations models in monetary economics." *Carnegie-Rochester Series on Public Policy*, 18, 1983, pp. 9–44.

Radford, R.A. "The economic organisation of a P.O.W. camp." *Economica*, 12, 1945, pp. 189–201.

Sephton, P.S. "Money in the production function revisited." *Applied Economics*, 20, 1988, pp. 853–60.

Part VIII

Money and financial institutions in growth theory

24 Monetary growth theory

This chapter assumes prior knowledge of the Solow growth model and of endogenous growth theory.

It is vital for monetary growth analysis to distinguish between the quantity of money as currency and balances in banks from the financial institutions and the payments mechanism, since these lead to different implications for the economy's growth.

This chapter also attempts to apply the notions behind endogenous technical change to financial intermediation. This is a comparatively new topic. Its discussion in this chapter is a first step and illustrative rather than definitive.

Growth theory does not provide definite conclusions on the significance of the role of money or financial intermediation to growth, but there does exist some empirical evidence supporting financial development as a contributor to the economy's growth.

Key concepts introduced in this chapter

- Commodity money
- Fiat money
- Inside money
- Steady-state rate of inflation
- Usage of money and the reduction of payments/transactions time in exchanges
- Financial services as an intermediate good
- Use of money in exchange as a mechanism for using the services of financial intermediation
- Invention and innovation in the financial sector
- Endogenous growth theory

The standard neoclassical (Solow, 1956) growth model in macroeconomics does not include money or a financial sector. This chapter will first modify this model in stages to incorporate money. Three types of money will be considered. The first one is a commodity money and the second is a costlessly supplied fiat money. The third type is inside money supplied by private financial intermediaries. The supply of inside money involves the use of labor and capital, so that in considering inside money we shift the focus from nominal or real balances to financial intermediation as an industry, with real balances representing tokens for the use of the services of the financial sector.

The economic literature on the importance of money and financial institutions to economic growth and development can be classified into several broad groups:

A. Studies that consider the contribution of money to output per capita in the economy to be zero or minimal.
B. Studies that maintain that financial development is an endogenous response to economic growth; the financial sector responds to the demand for financial services from other sectors by expanding the amount and variety of its services. Hence, causality runs only one way, from economic development to financial development.[1]
C. Studies that consider money and financial institutions to be vital to the prosperity of nations and therefore see these as making substantial contributions to output per capita and its growth. While this occurs at all stages of a nation's development, it is especially so in its early stage.
D. Further to (B), in the steady state, the contribution of money and financial institutions to output per capita can be only a level effect, or there could also be a growth effect on output per capita. Some studies, possibly a majority, advocate the former position of only level effects, while other studies assert both level and growth effects.

Approaches to money in growth theories

There are three main areas of differentiation in the approach to money in growth analyses. These involve:

1 The consideration of money in terms of its nominal quantity, and especially of the nominal quantity of fiat money,[2] versus the consideration of this nominal quantity as "claims" on the services of the financial sector, with the emphasis in the analysis of this sector being on its efficiency and innovation. The former view can be caricatured by treating money as a "veil," which affects only the appearance but not the magnitudes of the real variables, such as national output and its growth rate. The latter view treats the financial sector as critical to the efficiency and performance of all production and exchange in the economy and, therefore, critical to output and growth in the economy.[3]
2 The consideration of the role of fiat money and financial institutions in a static economy, or in one with technical change only in the production of commodities, versus its role in a dynamic context, with interactive innovation, growth and development of the financial, production and exchange sectors of the economy.
3 The consideration of the role of an unregulated, competitive and efficient financial sector, compared with the costs imposed on the economy by regulation, inefficiency and lack of competition in that sector. This issue is particularly relevant in discussions on why some countries develop faster than others even when they share the same knowledge of production techniques for commodities.

These differences lead to a dichotomy in the ways money is introduced into growth theory. On one side of this dichotomy is the introduction of money as only fiat money into the

1 Robinson (1952) provides an early example of this argument. Lucas (1988) called financial development to be a "sideshow" of the economy.
2 The overlapping generations models of Chapters 21 and 22 fall into this pattern.
3 The role of money in the production function in Chapter 3 falls into this pattern.

neoclassical growth theory, without an explicit consideration of the financial sector. This is treated in Sections 24.2 to 24.4 below. On the other side of the dichotomy is the focus on the financial sector as the user of inputs to provide services such as for the exchange of goods, the hiring and payments for inputs, mobilizing and allocating savings, the efficient funding of investment projects, etc., and often innovating in its provision of these services. The latter calls for consideration of the financial sector, including the organized markets for trades in bonds and stocks. There is no generally accepted model or even framework for the latter, so that only an introduction to this approach is presented in Sections 24.5 and 24.6.

Section 24.1 presents the basic model with one of the commodities serving as money. Sections 24.2 to 24.4 present in stages the modifications of the basic neoclassical growth model with exogenous technical change to incorporate fiat money. However, the main forms of money in the real-world economies are other than fiat money, and these – as well as many other liquid assets – are provided by financial intermediaries, which use labor and capital to provide monetary services. This role of financial institutions is the subject of Sections 24.5 to 24.9. Section 24.10 presents the empirical evidence on the importance of the financial sector to economic growth. Section 24.11 draws upon the arguments in the endogenous technical change literature to analyze the contribution of money and financial institutions to output per capita and to its growth, and Section 24.12 links investment, financial intermediation and economic development.

Stylized empirical facts on the importance of money in output growth

As we will see from the models developed in this chapter, the neoclassical monetary growth models imply that monetary economies can have higher or lower output than corresponding barter (non-monetary) economies. In order to judge among these implications and use selectivity in not pursuing clearly counter-factual possibilities, we need to establish some stylized empirical facts about growth in monetary economies. We state two of these as:

1 The least controversial stylized fact about the contribution of money in the economy is that monetary economies have higher output per capita and per worker than barter economies.
2 Another stylized fact is that, for a given capital–labor ratio, monetary economies have higher growth rates than barter ones, though whether this higher growth rate also holds in the steady state is more controversial.

The problem in interpreting and applying these statements is about the nature and domain of the *ceteris paribus* clause in them. Monetary economies over time develop production and exchange technologies which are drastically different from those of the barter economies which preceded them: monetary exchanges as against barter ones facilitate specialization in skills, commodity production, location, regional and international trade, etc. As an example of the drastic differences between monetary and barter economies, the modern large corporations could not really exist in a barter economy, nor could Internet-based exchange transactions. In this sense, the usage of money has not merely proved to be a veil covering the surface of the real economy, it has shaped and altered the structure of the economy and increased its size.

Therefore, a realistic evaluation of the historical benefits of money over the long run has to allow for the resulting changes in the structure of production and the economy. However, it is difficult to pin down exactly what parts of the changes in technology and exchange patterns

are due to the use of money versus those that could have equally easily arisen in non-monetary and more primitive financial economies. This is impossible to determine empirically. Given this difficulty and economists' limited analytical apparatus, the usual *ceteris paribus* clause – with its assumption of an unchanging technology even in the long run between barter and monetary economies, and between economies with more primitive financial versus more advanced financial intermediation – is often imposed in modeling and colors many of the results derived from monetary growth models. Great care must therefore be used in applying results based on such a *ceteris paribus* clause.

24.1 Commodity money, real balances and growth theory

This section presents the basic neoclassical monetary growth model using the production technology of the Solow (1956) model, with one of the commodities serving as a medium of payments and thereby acting as money. It is assumed that there are no costs of using this commodity money, that it is durable, and that it does not enter the utility and production functions. The overall scenario is really that of a barter economy just entering into multilateral exchanges by converting some of its output to usage as money, as against its usage in consumption or production.

Suppose that the economy has per capita output y of commodities, which constitutes its per capita disposable income in each period. Its per capita consumption c and per capita saving s, as in Solow (1956), are assumed to be given by:

$$\begin{aligned} c &= (1-\sigma)y \\ s &= \sigma y \end{aligned} \tag{1}$$

Part of this saving is used to increase the economy's real balances by m', with the remainder s_k allocated to increase its physical capital stock used in production, so that:

$$s_k = \sigma y - m' \tag{2}$$

where s_k is the change in per capita physical capital, m is the per capita real balances and m' is the change in m. σ is the marginal/average propensity to save out of income.

Assume that the demand for real balances of fiat money m^{d} is a proportion[4] λ of output y, so that:

$$m^{\mathrm{d}} = \lambda y \tag{3}$$

Assuming equilibrium in the money market through instantaneous adjustment of the price level,

$$m = m^{\mathrm{d}} = \lambda y \tag{4}$$

which implies that $m' = \lambda y'$ and $m'' = y''$, where $'$ indicates the change and $''$ designates the rate of growth. Note that if $\lambda = 0$ the system has no money and represents a barter economy.

4 This proportion can be taken as a constant or as dependent upon the opportunity cost of holding real balances. This cost equals the real rate of return on physical capital (equal to the real rate of interest) plus the expected rate of inflation.

Since $y'' = y'/y$, we have $y' = y''y$. Hence, the increase m' in real balances per capita is:

$$m' = \lambda y''y \tag{5}$$

Therefore, from (2) and (5),

$$s_k = \sigma y - \lambda y''y \tag{6}$$

Hence, the steady-state (SS) condition is:

$$k' = s_k - nk = 0 \tag{7}$$

where:

k = capital/labor (K/L) ratio
k' = change in k
s_k = per capita saving available for investment in physical capital ($= k'$)
n = growth rate of labor supply L
nk = capital per worker required to equip the increase in workers with the existing capital intensity.

With $y = f(k)$, where $f(k)$ is the production function with $f' > 0$ and $f'' < 0$, we get from (6) and (7) that:

$$f(k)[\sigma - \lambda y''] - nk = 0 \tag{8}$$

Therefore, the SS condition of the assumed economy is:

$$f(k)[\sigma - \lambda y''] = nk \tag{9}$$

Equation (8) is graphed in Figure 24.1, with k on the horizontal axis.[5] The curve for $f(k)[\sigma - \lambda y'']$ is shown as X'. This curve will be concave and lower than for $\sigma f(k)$, which is the curve X for the barter economy (with $m \equiv \lambda \equiv 0$). As Figure 24.1 shows, the SS value k^*_1 for the monetary economy will be less than k^*_0 for the barter economy. However, at k^*_1, k is constant, so y is also constant. Hence, the SS growth rates of capital K and output Y are given by:

$$K'' = Y'' = n$$

which are identical to the growth rates in the absence of money. Hence, the commodity-money economy and the barter economy will have the same SS growth rates (y'' and Y'') in this model but the SS values y^* and k^* will be *lower* for the monetary economy which has, by assumption, the same production function as the barter economy. Therefore, the per capita incomes and standards of living will be lower in the commodity-money economy. Further, the introduction of commodity money into the Solow model produces effects on the SS *level* but does not change the SS *growth rate*.

5 We have not labeled the vertical axis of this figure. Its units are output units. The concave curves measure saving per capita while the linear nk line measures investment requirements for new workers to maintain the existing capital/labor ratio.

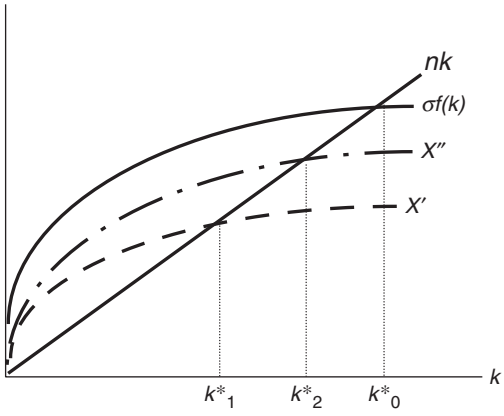


Figure 24.1

However, in the pre-steady state, the commodity-money economy does produce a lower rate of growth for any given capital/labor ratio as shown by the following argument. As in Solow (1956), the relationship between y'' and k'' for the production function $f(k)$ is:

$$y'' = \alpha' k''$$

where $\alpha' = y_k k/y$, so that α' is the output share of capital. Since k'' in the commodity-money economy is reduced by the need to provide some of the saving for increases in the real balances, k'' is less in the economy that uses a commodity as money than in the barter economy, and so is y''.

The implications of the preceding analysis are:

1 The SS per capita income y is lower in the commodity-money economy than in the barter economy.
2 The SS growth rates are the same.

Implication 1 clearly contradicts the stylized facts on money and growth specified at the beginning of this chapter.

24.2 Fiat balances in disposable income and growth

This section modifies the analysis of the preceding section by replacing commodity money by fiat money, which represents a common historical pattern of monetary evolution. The model presented in this section is based on the analysis of Levhari and Patinkin (1968), who had incorporated and extended the analysis of Tobin (1965) and Johnson (1967). Money in this basic monetary model is fiat money issued by the government and is thus "outside money." It is assumed that there are no costs of producing this fiat money and introducing it into the economy, nor of using and holding it. Newly created fiat money is assumed to be transferred gratuitously and directly to the public. Alternatively, the seigniorage from its creation is used by the government to buy goods that are then

provided free to the public, which would happen under a money-financed fiscal deficit. Under both these alternatives, the commodity income of the economy is not reduced by the introduction of fiat money into the economy. Hence, the disposable income of the public consists of the economy's output and the increase in real balances in the economy. That is,

$$y_d = y + m' \tag{10}$$

where y_d is disposable per capita real income and, as derived earlier, $m' = \lambda y'' y$. Hence, from (5) and (10),

$$y_d = y + \lambda y'' y \tag{10'}$$

Assuming as before that the average propensity to consume is constant at $(1 - \sigma)$,

$$c = (1 - \sigma)[y + \lambda y'' y] \tag{11}$$

Hence, per capita saving s out of per capita output y is:

$$\begin{aligned} s &= y - c \\ &= \sigma y - (1 - \sigma)\lambda y'' y \\ &= y[\sigma - (1 - \sigma)\lambda y''] \end{aligned} \tag{12}$$

Since fiat money does not use any commodities, all of s is used to increase the physical capital stock, so that $s_k = s$. Hence, the Solow growth equation becomes:

$$k' = f(k)[\sigma - (1 - \sigma)\lambda y''] - nk \tag{13}$$

Therefore, in the steady state,

$$f(k)[\sigma - (1 - \sigma)\lambda y''] = nk \tag{14}$$

The concave curve specified by the left side of (14) is shown as X'' in Figure 24.1. X'' will lie below X (that is, for the barter economy). Hence, in the presence of fiat money, X'' implies a lower SS value of k^*, at k^*_2, than k^*_0 in the barter economy. But, comparing (9) and (14), since $0 < \sigma < 1$, X'' (for the fiat-money case) will be above X' (for the commodity-money case), so that $k^*_2 > k^*_1$. This implies higher SS values of y^* and k^* for the fiat-money case than for the commodity-money case, so that the economy becomes better off in level terms by the switch to fiat money from commodity-money, even though it is still worse off than the barter case. Note that the SS growth rates $y^{*\prime\prime}$ and $k^{*\prime\prime}$ are identical for all cases. Hence, the introduction of fiat money into the economy, just as for the commodity-money, affects only the SS level but not the SS growth rate.

However, in the pre-SS equilibrium for the fiat money case as compared with the barter economy, the reduction in saving in the form of commodities due to the increase in the consumption of commodities reduces k'', which implies that y'' is also lower. That is, the

pre-SS growth rate is lower in the fiat money economy than in the barter economy. This reduction in the growth rate of output is less than for the commodity-money case.

Steady-state rate of inflation

We have so far considered only the steady-state condition for capital. Since we now have another asset, real balances, in the model, its SS condition is:

$$m' = 0$$

where $m = M/PL$, so that:

$$m' = m''m = (M'' - P'' - L'')m \tag{15}$$

Therefore, the SS condition is:

$$(M'' - P'' - L'')m = 0 \tag{16}$$

Let $M'' = \theta, P'' = \pi$ and $L'' = n$. Hence, for $m > 0$, (16) implies that:

$$\theta - \pi = n$$

and the SS rate of inflation π^* is given by:

$$\pi^* = \theta - n \tag{17}$$

which implies that the SS inflation rate will equal the difference between the money and the labor growth rates.

Predictions of the model

Among the predictions for the real world, the preceding models imply that:

1 SS output per capita and capital intensity in both the fiat-money and commodity-money economies would be *lower* than in the corresponding barter economy.
2 Growth rates will also be lower at given values of the capital/labor ratio outside the steady state.
3 SS growth rates will be the same.

These predictions indicate that the economy is better off without the use of money. This is counter-factual since rational economic agents will then prefer the barter economy to the money one and the barter economies will not evolve into monetary economies, which they have invariably done in all societies (Radford, 1945). Further, our stylized facts argue that monetary economies invariably have higher output per capita than barter economies. Therefore, the preceding model misses important elements of the role of money in the economy and must be modified in order to derive more realistic implications. This will be attempted in the next section by introducing real balances into the production function.

24.3 Real fiat balances in the static production function

We will take it as an established fact that the use of money in the economy allows the latter to produce more with given levels of capital and labor. To ensure this, there are two routes for the modification of the above model. One is to redefine output to include in it the *services* of money in exchange and production.[6] The other is to assume that these services are captured by putting real balances in the production function and assuming that real balances have a positive marginal product (Levhari and Patinkin, 1968). We will proceed along the latter route and follow the Levhari and Patinkin analysis.

To incorporate this suggestion into the fiat money model of the preceding section, redefine the production function of the representative firm as:

$$y = g(k, m^{\mathrm{f}}) \qquad g_k, g_m > 0; g_{kk}, g_{mm} < 0 \tag{18}$$

where m^{f} are the real balances held by the firm. With this modification, the SS condition corresponding to (14), in which increases in the real balances per capita increase disposable income and consumption but decrease saving, becomes:

$$g(k, m^{\mathrm{f}})[\sigma - (1-\sigma)\lambda y'] = nk \tag{19}$$

The left side of (19) is the saving available for investment in physical capital. By (18), for a given k, the presence of real balances increases the economy's output and therefore increases its saving. However, the transfer of seigniorage to the public increases disposable income and, through the income effect, increases consumption and reduces saving. The net effect of these opposing influences on the saving available for investment could go in either direction. Hence, if we were to show the left side of (19) by a curve X''' (not actually drawn) in Figure 24.1, X''' could lie above or below the curve σy for the barter economy. However, it would definitely lie above X' and X'', so that the SS value of k could be higher or lower than in the non-monetary economy, but definitely higher than if money were not in the production function.

Hence, the implications of this model are:

1 SS output per worker could be higher or lower for the monetary than for the non-monetary economy.
2 The SS growth rate would be the same, at n, for both economies.

However, we argued in the introduction to this chapter that a monetary economy produces substantially higher output than a corresponding non-monetary one with the same inputs. To obtain this result in equation (19) and Figure 24.1, we would have to add to our analysis the *ad hoc* postulate that the net effect of the use of money is an increase in saving for any value of k. This requires that the increase in saving due to the increase in production dominates the reduction in saving due to the creation of new money. However, there is no theoretical basis for making this assumption as a general case, so that we do not set out the analysis for such a shifts in production and saving.

6 If money is not costlessly supplied and costlessly used fiat money but needs labor and capital to generate its services, its use will increase output only if there is a net gain in output from the reallocation of some capital and labor to the production of monetary services.

24.4 Reformulation of the neoclassical model with money in the static production and utility functions

Steady state with money in the utility function (MIUF) and in the static production function (MIPF)

To determine the saving available for investment, we need first to derive the households' demand for real balances. One way of deriving it is to include real balances in the utility function. This is the money-in-the-utility function (MIUF) approach of Chapter 3, and was proposed by Sidrauski (1967) for monetary growth theory. The following model is based on his analysis. It assumes that:

$$U = U(c, m^{\mathrm{h}}) \tag{20}^{7}$$

where:

$U(.)$ = household's utility function
c = consumption per worker
m^{h} = real balances per worker held by households.

Sidrauski argued that, since the services of real balances yield utility, the imputed value of these services should be added to the output y of commodities and the seigniorage m' from money creation in order to calculate the household's total income. In equilibrium, the real value of the services of a unit of real balances will equal the market rate of interest r_m, which in turn equals the real rate of return r^{r} plus the (actual) rate of inflation π.[8] Households receive amounts y, m' and $r_m m^{\mathrm{h}}$ per period. Hence, the per capita disposable income y_d of households is now modified to:

$$y_d = y + m' + r_m m^{\mathrm{h}}$$

where $r_m = r^{\mathrm{r}} + \pi$, so that:

$$y_d = y + m' + (r^{\mathrm{r}} + \pi) m^{\mathrm{h}}$$

where $m'^{\mathrm{h}} = M^{\mathrm{h}}/PN$ and $m' = (\theta - \pi - n)\, m^{\mathrm{h}}$ as in (15). Hence,

$$y_d = y + (\theta - \pi - n) m^{\mathrm{h}} + (r^{\mathrm{r}} + \pi) m^{\mathrm{h}} \tag{21}$$

where y is the real income from commodities, $[(\theta - \pi - n) m^{\mathrm{h}}]$ is the increase in real balances so that it is the real value of seigniorage, and $[(r^{\mathrm{r}} + \pi) m^{\mathrm{h}}]$ is the real value of the services from holding money. Simplify (21) to:

$$y_d = y + (\theta + r^{\mathrm{r}} - n) m^{\mathrm{h}} \tag{22}$$

7 For a simpler analysis, we have used a timeless framework rather than the more elaborate intertemporal utility function of Sidrauski (1967).

8 We have assumed here that inflation is fully expected, so that the expected rate of inflation equals the actual rate. This is really an assumption of certainty or perfect foresight. Such an assumption is common in growth theory and stronger than that of rational expectations under uncertainty.

Households maximize (20) with respect to c and m^{h}, subject to (22). Assuming, as before, a constant average propensity to consume at $(1-\sigma)$,

$$c = y_d - s = (1-\sigma)\{y + (\theta + r^{\mathrm{r}} - n)m^{\mathrm{h}}\} \tag{23}$$

where c is per capita consumption. However, $[(r^{\mathrm{r}} + \pi)m]$ is spent on *using the services* of real balances and only the remainder – that is, $[c - (r^{\mathrm{r}} + \pi)m^{\mathrm{h}}]$ – is spent on commodity purchases. Per capita saving s_k available for investment in physical capital is:

$$s_k = y - \{c - (r^{\mathrm{r}} + \pi)m^{\mathrm{h}}\} \tag{24}$$

$$= y - [(1-\sigma)\{y + (\theta + r^{\mathrm{r}} - n)m\} - (r^{\mathrm{r}} + \pi)m^{\mathrm{h}}]$$

$$= \sigma y - (1-\sigma)(\theta + r^{\mathrm{r}} - n)m^{\mathrm{h}} - (r^{\mathrm{r}} + \pi)m^{\mathrm{h}} \tag{25}$$

$$= \sigma y - \{(1-\sigma)(\theta + r^{\mathrm{r}} - n) - (r^{\mathrm{r}} + \pi)\}m^{\mathrm{h}} \tag{26}$$

where the impact of real balances m^{h} on s_k is ambiguous: s_k exceeds σy if the term $\{.\}$ is negative but is less than σy if $\{.\}$ is positive. The former is more likely the higher θ is relative to π. Note that $\pi = \theta$ does not necessarily make s_k equal to or greater than σy.

In the steady state, $s_k = nk$, so that (26) implies the SS condition:

$$\sigma g(k, m^{\mathrm{f}}) - \{(1-\sigma)(\theta + r^{\mathrm{r}} - n) - (r^{\mathrm{r}} + \pi)\}m^{\mathrm{h}} = nk \tag{27}$$

Note that the firms' holdings of real balances are in the production function $g(k, m^{\mathrm{f}})$.

For the SS growth effects, since the steady state requires that $k' = 0$, the SS growth rates of capital and output in the present version of the model are still n. Hence, this introduction of money into the utility and production functions does not change the SS growth rate.

For the SS *level* effects, if we were to ignore the contribution of m^{f} in the production function, then, from the discussion following (26), s_k can increase or decrease as a consequence of the usage of real balances, so that the SS capital and output per worker in this monetary model could be higher or lower than in the corresponding barter model. To get a definitive increase in these variables for the monetary model, there must be large positive effects of the usage of money on output in the monetary economy. This points to the very significant role of m^{f} in the production function in (27).

For pre-SS analysis with values of k less than the SS value, the rates of growth for any given value of k are greater after monetization if the left side of (27) exceeds $\sigma y(k)$ (for the barter economy), but are lower in the converse case. The latter implication highlights once again the critical role of money in the technology of production.

Real balances and inflation in the steady state

For the demand for real balances by households in the MIUF approach, utility maximization yields the households' demand for real balances m^{hd} as:

$$m^{\mathrm{hd}} = m^{\mathrm{hd}}(r^{\mathrm{r}} + \pi, y_d) \tag{28}$$

For the holdings of real balances by firms in the MIPF approach, profit maximization by the firms implies that firms will be indifferent between holding capital or real balances if their

net returns are the same. The net return on capital is its marginal product $g_k(k, m)$ whereas the net return on real balances is its marginal product less the rate of inflation, which is the loss due to inflation from holding real balances. Hence, the portfolio equilibrium condition for firms is:

$$g_m(k, m^{\mathrm{f}}) - \pi = g_k(k, m^{\mathrm{f}}) \tag{29}$$

where m^{f} are the real balances held by firms. (29) yields the demand for real balances m^{fd} by firms as:

$$m^{\mathrm{fd}} = m^{\mathrm{fd}}(\pi, k) \tag{30}$$

Hence the equilibrium condition for real balances, with m as the supply of real balances per capita, is:

$$m = m^{\mathrm{d}} = m^{\mathrm{hd}}(r^{\mathrm{r}} + \pi) + m^{\mathrm{fd}}(\pi, k) \tag{31}$$

To derive the SS growth rates of real balances and inflation, note that the steady state requires that $m' = 0$. That is, as in (16) for the steady state,

$$m' = (M'' - P'' - L'')m = (\theta - \pi - n)m = 0 \tag{32}$$

which implies for $m > 0$ that:

$$\theta - \pi = n$$

so that the SS rate of inflation π^* is again given by:

$$\pi^* = \theta - n \tag{33}$$

Hence, in the steady state, the real balances (M/P) grow at the growth rate n of labor whereas the SS inflation rate π^* becomes constant at $(\theta - n)$. Further, with the SS values of k and m^{f} being constant in the production function, the SS real rate of interest $r^{\mathrm{r}*}(= f_k(k, m^{\mathrm{f}}))$ would also be a constant. Furthermore, the SS nominal rate of interest r would be constant at $(r^{\mathrm{r}} + \pi^*)$.

Increase in the labor force participation rate in production due to a monetary economy

Next section's analysis of how the use of money contributes to production shows that one of its major effects is to increase very significantly the participation rate of the labor force in production. Therefore, for a proper comparison of the monetary economy with a barter system, the preceding models need to be revised to incorporate a drastic increase in the labor force participation rate in production. For a given population, a higher labor force participation rate in production increases the standard of living, which is output per capita (not per worker), so that a monetary economy would have much higher standards of living than a barter one. However, assuming that the labor force participation rate in production

becomes constant in the steady state, its incorporation into the model would not alter the conclusions on the steady-state growth rates.

Note that shifts in the efficiency of the payments system in monetary economies are likely to release more of the labor force for use in production, so that such shifts will produce further increases in the standard of living.

24.5 Why and how does money contribute to per capita output and its growth rate?

The failure of the above modifications of the neoclassical growth model to show an unambiguous increase in output per worker for the monetary economy over that in its barter counterpart, let alone an increase in its growth rate, is certainly disappointing. We consider this failure as evidence of this model's inappropriateness for the monetary economy because of its failure to provide a sufficiently realistic role for money in the economy.

We illustrate our views on the role of money in the economy by the following analogy. Consider, for example, the use of artificial (non-human and non-animal) energy, especially electrical, in the economy. There is a fundamental difference between the production structures and prosperity of economies using only human and animal energy in production and those heavily reliant on artificial energy. Yet the only way the latter's impact on the economy can be shown in models is through a shift or periodic shifts in the economy's commodity production function due to the innovation and evolution of the forms of artificial energy usage. As an intermediate good, the impact of electricity consists of two opposing forces: electricity generation uses some labor and capital, thereby reducing their amounts left over for use in the production of final goods, while the usage of electricity as an intermediate good increases the output of the latter. Therefore, without knowing the actual shift in the production function of final output due to the use of electricity, it is not possible to determine a priori whether the switch to electricity will increase or decrease final output, and especially whether it will do so at the margin of existing usage.

From another perspective, the optimal production of electricity, and hence the optimal amount of the diversion of labor and capital to it, would be that which maximizes the output of the final commodities. Economic theory implies that in perfect competition, without externalities etc., the actual amount of electricity produced and used will be the optimal one, so that a competitive economy with a positive usage of electricity will clearly constitute evidence that the conversion to electricity did increase output per capita. However, note that this is an argument for a positive level effect, not one for a growth effect, of the invention and widespread usage of an intermediate good. There may also be a growth effect, but to show such an effect will require additional special arguments – for example, such as evidence for greater technical change in this industry relative to the average rate for the economy.

In our view, money, just like artificial energy (e.g. electricity), is an intermediate good in consumption and production. As in the case of energy, money's contribution to output per capita and output growth cannot be assessed without specifying the shifts and evolution of the economy's commodity production function under the impact of the innovation and evolution of money and financial intermediation. One implication of this argument is that the commodity production function is different in the pre- and post-versions of the innovation *and* widespread use of money, as it is in the pre- and post-versions of the innovation and extensive production of artificial energy. Hence, in our view, it is a mistake for growth theory to assume that the commodity production function in the barter economy would be the same

as in the monetary economy, with the result that the monetary growth theory based on such an assumption is of doubtful value. Another implication is that monetary growth theory must not only include real balances in the commodity production function but must also explicitly model the role of financial intermediation in the economy and that of innovations in the payments system. This is attempted in the next section.

Further, the output of financial intermediation (FI), just as that of energy, need not be at the optimal level for commodity production. Note that the actual production of these sectors in any given economy can be non-optimal: either held below or induced to be above this optimal amount, depending upon the structure of the industry, externalities and rigidities, governmental policies and controls, the availability of investment funds, etc. However, in terms of the preceding monetary growth model, this model, in common with most such models, implicitly assumes that the financial sector, like any other sector, functions in a competitive and unregulated manner, and that the market for its services is continuously in equilibrium through price adjustments. But many countries regulate their financial sectors closely, sometimes over long periods, with some setting interest rates or credit allocations and imposing other restrictions. A common question asked in such cases – especially in the context of developing countries – is whether the liberalization and expansion of the financial sector would increase the standard of living and the growth rate of the economy. This is a different question, with a different perspective on money in the economy, from that considered so far in this chapter. We examine this issue as part of the broad consideration of FI in the next section.

24.6 How does the use of money change the labor supplied for production?

Chapters 3 and 23 introduced the concept of payments/transactions time in the individual's utility function, so that the worker has to budget for the additional transactions time required to purchase the commodities bought with the income from the labor he has supplied. The required transactions time is an ancillary cost of labor supply (to earn income to pay for the shopping expenditures), so that the labor supply is a decreasing function of its ancillary transactions time. This transactions time is substantially less in a monetary economy than in a barter economy without the double coincidence of wants.[9] In fact, the transition from a barter to a monetary economy represents a drastic reduction in the amount of labor used in exchange, accompanied by a large increase in the labor supply for use both in production for the market by firms and within the household. This is illustrated quite well by Exhibit A ("An experiment in reality") in Chapter 23 on the breakdown of monetary arrangements after the collapse of the Soviet Union. Following this breakdown, labor had to drastically increase the amount of time devoted to exchanges (for example, on roadside stands) at the cost of the time supplied to market and household production.

Therefore, going from a barter to a monetary economy results in a very drastic increase in the labor supply devoted to production, at the expense of shopping and selling time.

9 In a barter economy, a good part of the available time of a producer would have to be devoted to the direct sale to the ultimate consumers of his output. Imagine a farmer going around the countryside and towns with his produce (wheat, vegetables, flowers, etc.) to find consumers for it, and compare the time it takes him to produce one unit of the good versus the time needed to sell it, especially if he wants to get in exchange something that he is going to use himself.

This increase in the labor supply, in turn, implies a substantial increase in the economy's output. There would also be additional effects, such as those of allowing specialization, increasing the size of the market, etc., which occur with the monetization of the economy.

To summarize, the production and labor supply functions in functioning monetary economies are likely to be significantly different from their counterparts in barter economies. Whether these shifts affect only the steady-state level of output or its growth rate as well still remains to be discussed.

24.7 Distinction between inside and outside money

Inside money represents monetary assets that are also a liability of the individuals and firms, such as the banks, within the private sector. If the accounting procedure of summing over the assets and the liabilities of the private sector to calculate its net worth is followed, the assets represented by inside money are offset exactly by the liabilities represented by it, so that the net worth of the private sector does not include inside money. Demand and savings deposits are assets of the households and firms that hold them but are a liability of the banks in which they are held. Therefore, they are not part of the accounting calculation of the net worth of the private sector since banks are units within the private sector.[10]

Outside money is an asset of the private sector but a liability of the government sector, including the central bank. As such, the accounting calculation of the net worth of the private sector will include outside money. The monetary base is an asset held by the private sector, including the banks, but is not a liability of any units of the private sector. Therefore, the monetary base is part of the net worth of the private sector. As a liability of the government, it represents outside money.

Another distinction between inside and outside money is that the former requires the usage of capital and labor while the provision of the monetary base, which is fiat money, does not. This distinction – rather than whether one of them is part of the private sector's net worth and the other is not – is the one relevant to the rest of this chapter.

24.8 Financial intermediation (FI) in the growth and development processes

The preceding mode of specifying monetary growth theory was based on the introduction of the real balances of fiat money at various points in neoclassical growth theory. These points were:

1 the definition of disposable income;
2 the allocation of saving to capital and to real balances;
3 the insertion of real balances into the utility and production functions.

The concept of money in these models was that of fiat money, costlessly created without using any part of the economy's labor and capital in its production or usage. However, fiat money in circulation – which constitutes currency (notes and coins) in the hands of consumers and firms other than financial intermediaries – is a very small part of the actual moneyness

10 However, note that there are disputes in the literature on whether inside money should be included in the economic calculation of net worth, and how to calculate this economic net worth.

or liquidity in the modern economy. Further, even the usage of currency imposes costs on the users; it has to be acquired, usually from financial intermediaries (by households) or from customers (by firms), with some expenditure of labor time and capital, and its usage also involves labor and capital in the acts of safe keeping, counting and handing over to others. These costs are sufficiently significant for both households and firms to prefer various types of monies created by the private sector, even though these require the expenditure of labor and capital both for their creation and usage and, therefore, impose a user cost.

While the usage of real balances diverts some of the economy's savings to them, their excessive regulation can impose even higher costs on the consumption and production sectors of the economy. The basic reason for this is that it is much more efficient to consume or produce if all households and firms operate with money (as against a barter economy) and if the households and firms have adequate balances to perform their transactions efficiently. This argument adduces the benefits from the use of money to be akin to the benefits, say, from the use of electricity (or of automobiles). For all of these, when no one else in the economy uses them, the costs to a single firm or household using them are very high, and the benefits are very small, since both electrical stations and car production plants and roads would then be uneconomic to build and would not exist. There are, therefore, considerable externalities of economy-wide usage in these cases.

Further, considerable efficiencies are to be gained from the competitive creation of alternatives to fiat money. These arguments suggest that the focus of the appropriate treatment of money should be on the provision of, and access to, the services of financial intermediation (FI), along with the consideration of the labor and capital used in the FI sector. Banks are part of this sector, as are near-banks, investment brokers, insurance companies, pension funds, etc., with each providing a variety of differentiated products.

24.9 The financial system

The financial system consists of:

1 financial instruments;
2 financial markets;
3 the central bank;
4 accounting and legal structures regulating the financial instruments, institutions and markets.

Financial instruments can be divided into several categories, of which the ones most relevant to our discussion on growth are settlement or payments instruments such as currency, checks, debit and credit cards that serve as media of payments, and investment instruments such as shares, bonds and futures. Settlement instruments reduce transactions costs and facilitate specialization in production and exchange. Reductions in transactions costs and the promotion of such specialization occur not only in the switch from a barter economy to a monetary one, but also with further innovations in the payments instruments, so that the innovations in payments instruments continually promote the growth of output.

Investment intermediation occurs in the collection and channeling of funds, originating in savings and other ways, to investment. In a perfectly competitive market with full and free access to information and full enforcement of contracts, loans will be made at rates of interest that fully reflect their individual risks and will vary with the default risk of the borrower.

However, as Chapter 16 on credit argued, market imperfections are prevalent in credit markets. These arise because of adverse selection, moral hazard and monitoring and agency costs, which create credit rationing by both quantity and price (i.e. the interest rate), as well as the need for collateral in loans. The evolution of bond markets reduces their impact and enables a substantially larger amount of external borrowing than in their absence. The availability of adequate funds through both banks and financial markets is essential to the achievement and growth of modern levels of economic activity.

Banks, other financial intermediaries and financial development

Financial development can be measured by the number and variety of financial intermediaries, the size and sophistication of the markets for bonds and stocks, and the efficiency of the rules, regulations and practices governing the financial practices of firms in the economy. Meltzer (1969) provides a review of the earlier views linking money, financial intermediation and growth. Greenwood and Smith (1997) model the links between financial development and economic growth; Diamond (1997) provides a model of financial sector development, with changes in the structure and market share of banks versus financial markets. The empirical study by Rajan and Zingales (1998) reports that financial development enhances long-run growth.

Although commercial banks are a cornerstone of the financial sector, they are only one part of it. This sector also includes other institutions such as the stock and bond markets, brokerage firms, pension funds, insurance companies, mutual funds, etc.[11] These institutions and markets are governed and supported by the rules, regulations and practices governing the financial system. These include the accounting and disclosure requirements, monitoring, certification and the public dissemination of information on firms, including financial intermediaries, as well as measures against fraud, inside trading and so on. These factors affect the extent to which firms can raise capital from external sources and, in a competitive system, reduce the cost of such capital, whether directly in loans from banks and other financial institutions or through the public offerings of shares and bonds. Well-developed and efficient financial markets allow firms to rely on external sources of financing their investments and to do so in a competitive manner. For savers, the high degree of liquidity of the investment in these markets is more attractive than the alternatives of hoarding and private illiquid investments.

One can, therefore, distinguish between the contributions of banking and of the rest of the financial system to the growth of output. However, this is rarely done at the general theoretical level and we will not do so in this chapter. However, this contribution can be separated in empirical studies, as reported later in this chapter.

Real balances as tokens for the use of banking services

From the perspective of financial intermediation, the real balances of inside money may be viewed as tokens for the use of banking services. But the holdings and use of these tokens

11 Banking activity usually becomes an increasingly smaller proportion of the financial sector as financial and economic development proceeds, so that it is relatively smaller in the richer countries. Although both the bank and non-bank sectors are larger and more active in rich countries, stock markets become relatively more active and efficient at higher levels of development (Demirguc-Kunt and Levine, 2001).

by the non-bank public require the banks to devote labor and capital, as can be witnessed by walking into any bank building (representing the use of physical capital) and dealing with the bank employees (representing the use of labor). Banks and other financial intermediaries can therefore be viewed as firms producing the services provided to owners of real balances, with these services being in the nature of intermediate rather than final goods, and charging the buyers for such services. Real balances are merely tokens allowing access to banking services.

Basics of incorporating the financial sector into production in the economy

Therefore, an important approach to the contribution of FI to growth is to focus on the FI industry as a provider of services to consumers and to other industries. As noted above, these services are produced with the help of labor and capital, and sold in the market for a price. This picture differs substantially from the picture of a costlessly provided and costlessly used fiat money in the neoclassical growth theory of the preceding sections. We will now deviate from this theory and posit that while money and other financial assets are provided by either the government or the private sector, they do not yield their services except through FI, which involves their production at a cost. These services are the products of financial intermediaries, who charge the users for such services.

Designate both the financial sector and its real output per worker by z, with Z standing for its total output. Note that even if the financial sector is defined as being synonymous with banks, its output will be different from the amount of real balances in the economy even when these are defined broadly to include all deposits with banks.

A simple neoclassical-type production function for the financial sector can be specified as:

$$z = z(k_z) \qquad z_k > 0, z_{kk} < 0 \tag{34}$$

where:

z = real output per worker of the FI sector
k_z = capital/labor ratio of the FI sector.

The neoclassical-type production function for the output of final commodities (excluding the services of the financial sector) can similarly be given the form:

$$x = x(k_x, z) \qquad x_k, x_z > 0; x_{kk}, x_{zz} < 0 \tag{35}$$

where:

x = real output per worker of final commodities
k_x = capital/labor ratio in the commodity sector.

We will add the plausible assumption that k_x and z are "co-operative" in the sense that $x_{kz} > 0$ and $x_{zk} > 0$, so that an increase in financial services z increases the marginal product of capital, and vice versa. However, they are still substitutes in producing any given value of x.

In the real world, some of the output of the financial sector will be used as intermediate good in consumption and some in commodity production. As a simplification, assume that all of the FI services produced in the economy are used by the firms in the production of commodities so that all of the financial sector's output is used only as an intermediate good in production, and none is to be counted in the economy's final output. This final output is designated as Y, which equals $L_x x$, where L_x is the amount of labor employed in the production of commodities and x is the output of commodities per worker.

The nature of z in the commodity production function needs to be clarified. (34) specifies the composite output z of financial services without specifying the actual assets that will provide it at least cost. Within this composite output, the production of different types of FI assets, with quantities z_i, would depend upon the technology of FI, the demand for z_i and the relative costs of producing z_i. For example, over time, the relative amounts of currency or demand deposits or savings deposits, etc., change with the amounts of output and capital per worker, depending upon the characteristics and relative costs of these assets, and upon the techniques of production such as the use of human tellers versus automatic teller machines.[12] In this sense, the services of FI are like those of the transport or the energy sectors, whose demand increases with production in the economy but where the particular products that are bought to provide the required service can vary with the technology, the level of production, the capital/labor ratio and the costs of inputs.

The comparison of FI with the energy and transport sectors is also apt in another way. Limiting the production of the latter by regulation or through the introduction of inefficiencies in their production affects the output of virtually all the sectors in the economy, as well as forcing a change in the structure of the economy. The net effect is likely to be a change in the composition of aggregate output as well as a reduction in it. The same pattern of effects also applies to administratively introduced inefficiencies in FI.[13] Conversely, an increase in the efficiency of FI causes it to provide its services to other sectors in greater variety and at lower cost, thereby inducing an increase in their production, so that the aggregate output in the economy would increase.

The economy benefits from FI if its existence – or, at the margin, its increase – for given amounts of labor and capital in the economy increases the aggregate output Y of commodities. This need not always be so since the output of the FI sector plays two opposing roles in the economy: from (34), it uses up some capital and labor in the production of the intermediate good but, as specified by (35), the use of this good in commodity production enhances the productivity of labor and capital in the latter. Hence, for a given level of labor and capital in the economy, the net effect of the expansion in the FI sector on the commodity output could go either way. However, economic agents are free to use financial services or not to use them, and would use them only if it would be to their advantage. Further, as mentioned in the introduction to this chapter, the historical experience is that monetary economies generate much larger output per capita than corresponding barter economies. Therefore, a plausible assumption for long-run analysis is that an increase in the output of the financial sector in an efficient market economy causes a net increase in the economy's commodity output in the observed range of production of FI.[14] We will therefore assume that, in the long run, $\partial y/\partial z > 0$ and $\partial^2 y/\partial z^2 < 0$. Under these assumptions, an increase in the output of the financial sector would cause an increase in the economy's per capita output.

Another relevant question in the context of growth theory is whether such an increase in the efficiency and output of the FI sector also causes an increase in the *growth rate* of the

12 For instance, currency seems to be the most convenient and cost-effective asset for transactions at very low levels of output and capital per worker, yielding to demand deposits as the latter increase.

13 To continue with our analogy, restricting all transactions in the economy to the use of only currency and demand deposits could be akin to restricting all transport in the economy to the use of only one kind of car and one kind of lorry. Both would introduce considerable inefficiencies in transport and considerably reduce the output of the other sectors of the economy.

14 However, this need not be true in the short run.

economy and, further, whether it increases the *SS growth rate* of the economy. There is no consensus in the economics profession on these issues, though historical experience suggests that an efficient and growing FI sector does seem to be a requisite for fast-growing economies.

A simple modification of (34) to encompass technical change in the production and financial sectors is:

$$z = z(\beta(t)k_z) \qquad z_k > 0,\ z_{kk} < 0 \tag{36}$$

where $\beta(t)$ measures the efficiency of capital and labor in the production of financial services.

The corresponding simple modification of (35) to encompass technical change in the production and financial sectors is:

$$x = x(a(t)k_x, b(t)z) \qquad x_k, x_z > 0;\ x_{kk}, x_{zz} < 0 \tag{37}$$

where $a(t)$ and $b(t)$ represent the efficiency of usage of the capital per worker and of the financial sector's output in the production of commodities. Changes in $b(t)$ can accrue because of innovations in the production sector or in the financial sector since the determinants of $b(t)$ include the diversity, liquidity and other characteristics of the latter's output. The financial development of the economy results in increases in either $\beta(t)$ or $b(t)$, or both.

The dispute on the role of financial development can now be framed as follows: are $\beta(t)$ and $b(t)$ caused wholly by $a(t)$ or can they occur independently, to some extent at least? In the former case, financial development is wholly induced by technical change in the production sectors; in the latter, the independent elements of financial sector growth can induce greater output in the production sector than would occur otherwise. While both of these are theoretical possibilities, the dispute on the contribution of financial development to the economy's growth revolves around real-world experience over long periods and can only be settled by empirical studies.

Equations (36) and (37) implicitly assume that technical change is exogenous and merely a function of time. But such change could also have endogenous elements. We consider this possibility in a subsequent section.

24.10 Empirical evidence on the importance of money and the financial sector to growth

We briefly and selectively consider some evidence on this issue from economic history and econometric studies. Corresponding to the pattern of the theoretical analyses in this chapter, we need to examine the empirical evidence on two separate issues: the effect of increases in the quantity of money for a given financial structure, and the effects of shifts in the financial structure. On the contribution of the financial structure, note that banks provide different services from those provided by stock markets, and that each can have independent effects on economic growth. More broadly, the banking system is only a subset of a developed financial sector whose other segments are also vital to growth in the modern economy.

Another point worth noting is that the contribution of the financial sector to growth rates occurs at least partly through the dynamics of the economy's industrial structure, especially in the external financing of new firms and new products. This ties in very well with the emphasis on innovation and change in the new theories of endogenous technical change.

Empirical evidence on the quantity of money, inflation and growth

The effect of changes in the quantity of money on growth can be examined in one of two ways. It can be estimated directly by using the data on the money supply, or indirectly, since the growth rate of the money supply is usually related to the rate of inflation, by examining the effect of inflation on the growth rate of output. The theoretical findings on these can be summarized as:

1 The quantity of money is (more or less) neutral in the long run.
2 The long-run output growth rate is also independent of both the money supply growth rate and the inflation rate.
3 Any deviations from neutrality emanating from the analyses in Sections 24.1 to 24.3 are minor (of a second or third order of importance, which may not be visible in the data) and may vary among different periods or countries.
4 Some of the inflation–growth literature claims that high inflation rates are inimical to growth because they introduce distortions and inefficiencies in the economy. This literature further claims that price stability enhances economic certainty and promotes investment, which is conducive to growth.[15] Hence, the output growth rates will be negatively related to high inflation rates.[16]

Among the studies which report that the growth rate is independent of the rate of monetary growth and the inflation rate are those by Bruno and Easterly (1996), Lucas (1996) and McCandless and Weber (1995). For 100 countries, Lucas plotted the bivariate relationship between 30-year averages of the output growth rate and the M2 growth rate and argued that the plots showed the former as being independent of the latter.[17] His conclusion was that, in the long run, money is neutral for the output growth rate.

McCandless and Weber (1995) used time-series data for 110 countries and 30-year geometric average rates of growth. Their finding was that there was no correlation between inflation and output growth. Further, in general, there was no correlation between the growth rate of money – measured by M0, M1 and M2 – and real output.[18] The exception to this was a subsample of OECD countries, with a positive correlation between these variables.

While some empirical studies have reported some sort of a correlation – some reporting a negative one[19] and others a positive one[20] – between the growth rate and the inflation rate (or the money growth rate), the estimated relationship is often weak and usually not robust to changes in the data set or the inclusion of other likely determinants of growth.[21] As a result,

15 However, this effect is not considered to be large.
16 It should also be kept in mind that the inflation rate and the growth rate of output are negatively related by virtue of the quantity equation, so that an observed negative relationship between them cannot be used as *prima facie* evidence of a causal link from inflation to growth.
17 Lucas further showed that the correlation – for 110 countries with 30-year averages of the data – between M2 growth and the rate of inflation is 0.95.
18 The correlation between the growth rate of money supply and inflation was "almost unity."
19 For example, Barro (1996) reported a negative and statistically significant relationship between the output growth rate and the inflation rate. The elasticity was –0.025, so that disinflation of 10 percent would increase the growth rate by 0.25 percent.
20 For example, McCandless and Weber (1995) reported a weak positive relationship for the OECD countries. For inflation rates below 40 percent, Burno and Easterly (1996) did not find a relationship between output growth and inflation. At higher inflation rates, there was a negative relationship between these variables.
21 Levine and Renelt (1992).

a vague sort of consensus continues to exist in monetary economics that output growth is independent of the inflation rate and the money growth rate.

Empirical evidence on the role of the financial sector in growth: some conclusions drawn from economic history

One form of the evidence on the relationship between the financial sector and economic growth comes from the economic history of the early stages of industrialization in countries noted for industrial innovation. Cameron's (1967) assessment on this was that:

> Financial innovation, after all, is not so very different from technical innovation. The former is frequently necessary for realization of the latter.
>
> While it is rare for banks to finance directly a period of experimentation with a completely novel production technique by a new, inexperienced businessman or inventor … it is quite common for bankers to finance the expansion of firms that have already introduced successful innovations, and also to finance the adoption of the innovation by imitators.
>
> (Cameron, 1967, pp. 12–13).

The assessment by Schumpeter, an early proponent of the importance of technological change – in current terminology, of endogenous technical change – to development, was:

> The essential function of credit … consists in enabling the entrepreneur to withdraw the producer's goods which he needs from their previous employments, by exercising a demand for them, and thereby to force the economic system into new channels.
>
> (Schumpeter, 1934, p. 106).

These quotes point to the critical role of the financial sector in the innovation and restructuring processes which are elements of any technical change, including the endogenous type, and of economic development.[22]

Empirical evidence on the role of the financial sector in growth: recent empirical evidence

Support for a finance-led-growth hypothesis, deduced by Cameron and Schumpeter from economic history relating to periods of early commercial and industrial development, is provided by the empirical study by Rousseau (2003), who studied microeconomic links, including case studies, between financial development and growth in Amsterdam (1640–1794), England (1720–1850), the USA (1790–1850) and Meiji Japan (1880–1913). This study reports that the availability of finance mattered for resource mobilization in periods of rapid economic development because banks and other financial intermediaries promoted investment and commerce by generating information, accumulating funds, facilitating payments and providing working capital. In each case, financial institutions arose to facilitate external financing of trade and industry, making the financial constraints on the economic

22 As against these claims, Lucas (1988) claimed that economists "badly over-stress" the role of the financial system. Other economists have argued that the financial sector responds passively to growth in the production sector.

expansion of firms less restrictive. Rousseau also reports cross-section regressions on data, averaged over five or ten years for the periods 1850–1929 and 1850–1997 for 17 countries bordering the Atlantic Ocean (the "Atlantic economies"), and finds that the change in financial depth was more significant in the earlier development period (1850–1929) than for the "mature" period. He considers this finding to be consistent with that in King and Levine (1993) whose sample of about 80 countries included many in their early development phases. Hence, the stage of development matters for causality from financial development to growth.

On the links between inflation, financial regulation and economic growth, Chari *et al*. (1995), from a cross-section analysis of 88 countries, find that inflation by itself has negligible effects on the growth rate. However, there is a high correlation between the rate of inflation and the required reserve ratios, and there is a negative – and highly non-linear – relationship between these reserve ratios and the growth rates. They conclude that financial regulations and the interaction of inflation with regulation have substantial negative effects on growth. This evidence of the detrimental effects of financial regulation has the corollary that a more efficient and developed financial system is conducive to growth.

Support for the links between financial development and growth, drawn from economic history, is provided by an empirical study. Rajan and Zingales (1998) report from a cross-section econometric analysis of countries that industries that depend relatively more on external financing develop faster in countries with more developed financial sectors. The need for external financing varies among industries and depends on factors such as the initial project scale, the gestation period, the internal generation of funds through retained earnings and the amounts needed for further investment. Financial development reduces the cost of external financing and therefore promotes the growth of existing firms and industries dependent on such financing. It also promotes the establishment of new firms and industries, which usually account for a large part of new ideas, innovations and breakthroughs.[23] Financial development also reduces financial market imperfections, which usually favor internal financing and the growth of existing firms relative to that of new firms.[24] Therefore, the lower cost and easier access to external finance provides a mechanism through which financial development influences change and growth in the economy and, in the authors' view, is evidence of causality running from financial development to growth. They also argued that, at the country level, a country with a more developed financial market has a comparative advantage over others.

Levine and Zervos (1998) proceeded further and differentiated between banking development and stock market liquidity, where the latter was measured by "turnover," defined as the values of trades of domestic shares on domestic stock exchanges divided by the value of listed domestic shares, or by "value traded," which was defined as the value of the above trades divided by the domestic GDP.[25] Stock market liquidity therefore refers to the ease with which domestic equities can be traded. Banking development was measured by the bank loans to private enterprises divided by GDP. Levine and Zervos reported from a cross-section

23 In some cases, as much as two-thirds of the total growth of output in the economy came from new firms, so that their number and development was a major element of growth. Rajan and Zingales (1998) showed that the effect of financial development on the growth rate of new firms was almost double that for pre-existing firms.

24 Financial development also promotes higher accounting standards, which reduce the costs of obtaining external finance.

25 Levine and Zervos (1998) distinguished this variable from the size (in terms of the ratio of the value of the listed shares to GDP) of the stock market. This size did not prove to be a reliable predictor of growth. These authors also considered measures of international integration.

country analysis of 47 countries over the 1976–93 period that the *initial* levels of both banking development and stock market liquidity were independently and positively related to *subsequent* growth rates, capital accumulation and productivity improvement.[26] The effects were very significant and indicated causality from financial development to economic growth. Further, Levine and Zervos reported that the main impact of these variables on growth was through productivity increases rather than through capital accumulation (Levine and Zervos, 1998; Temple, 1999). On the latter aspect, Levine and Zervos did not find a significant link between private saving and either of their financial development variables. Nor did domestic private saving prove to have a strong relationship with international financial linkages.

Recent empirical studies tend to proxy financial development by a number of variables. Among such variables are the liquid liabilities of financial intermediaries and measures of credit extended by them, usually as ratios of GDP. The proxies for growth tend to include the growth rates of GDP per capita, capital stock per capita and total factor productivity. The overall findings of King and Levine (1993), based on the long-run average data for such variables, indicated a close correlation between the average level of finance and per capita GDP growth for virtually all of the 80 countries in their study. Further, financial development led per capita GDP growth, substantiating their argument that financial development promotes growth by increasing capital accumulation and improving its efficiency in production.[27] Demetriades and Hussein (1996) use the more appropriate stationarity tests and cointegration analysis to investigate whether a stable, long-run relationship exists between real GDP per capita and their proxies for financial development. These proxies are the ratio of bank deposit liabilities to nominal GDP and the ratio of bank claims on the private sector to nominal GDP. Their study focuses on 16 "not highly developed" countries. They find that there is little or no evidence that causality is unidirectional from finance to GDP growth. Furthermore, the relationship is country-specific.

Macri and Sinha (2001) difference non-stationary variables until they are stationary, and use OLS regression – not cointegration – analyses and multivariate causality tests. A wide array of variables, including outside money, is used to capture financial development. They report a positive and significant relationship between the income and financial variables for only half the Asian economies in their sample of eight countries. Causality results are quite mixed: some economies exhibit causality from finance to growth, others from growth to finance, and yet others suggest bi-directionality. Generally, therefore, the findings seem to corroborate those of Demetriades and Hussein (1996).

Recent empirical studies have sought to go beyond the impact of bank intermediation on economic growth to include the impact of the financial sector beyond banks. In this vein, Rajan and Zingales (1998), Levine and Zervos (1998), Levine (2003) and Beck and Levine (2002) have all sought to capture the impact of stock market liquidity on economic growth.

26 Their finding was that the significant variable for growth was stock market liquidity (i.e. ability to trade stocks) rather than stock market size (i.e. how many companies are listed). One reason for this is that higher liquidity increases the willingness of investors to fund new long term business ventures where there is a high risk and absence of dividends for many years but investors can sell their holdings, if they wish to do so, without incurring large costs.

27 However, their cross-country OLS did not test for the stationarity of the data series or use cointegration techniques for deriving the long-run relationship. Further, their cross-section analysis implicitly assumes that the finance–growth link is the same for all countries (or groups of countries) on average, when, as argued above and as pointed out by Demetriades and Hussein (1996), the financial sector could lead the "real" sectors in some countries while lagging in others.

Rajan and Zingales find that industries with high needs for outside financing develop fastest in countries with well-developed financial systems, including stock markets. Levine and Zervos proxy "stock market liquidity" by the "turnover ratio," defined as the ratio of the value of trades in domestic equities on domestic stock markets to GDP. Their conclusion is that high-income countries tend to have higher levels of stock market liquidity than developing countries, and that increased liquidity spurs economic growth. This view is further explored and supported by Levine (2003).[28]

To conclude, causality from financial development to economic growth is more apparent in the early stages of development. It is also likely to be more evident from historical and microeconomic case studies than from macroeconomic data. Further, this causality is more evidently related to external sources of funding for firms than to the quantity of money in the economy. Whether in the early or late stages of growth, the implementation of new technologies in production usually requires borrowed funds, for which banks often play a more significant role than financial markets in earlier stages of economic and financial development, while financial markets play a greater role in developed economies. Note that it was argued earlier in this chapter that there is likely to be considerable contemporaneous causality in development between the financial sector and the rest of the economy. Granger causality tests rely on leads and lags to judge causality, so that they do not disclose the extent of contemporaneous causality. However, contemporaneous causality is akin to two-way causality rather than to one-way causality, so that two-way causality is more likely to be found in the presence of contemporaneous causality.

24.11 A simplified growth model of endogenous technical change involving the financial sector[29]

One reason why an unregulated and efficient financial sector promotes growth in the economy as a *continuing process* can be seen from the endogenous growth theories. In these, endogenous increases in knowledge and innovation are the vehicle for increases in the growth rate and for a positive growth rate of per capita output even in the steady state.

We now bring in the arguments on endogenous growth from the growth literature and adapt them to take account of our arguments in the preceding section. To incorporate the role of the financial sector in the economy, modify the corresponding commodity production function to the form:

$$x_{ij} = [A^j {k_{xij}}^{\alpha'} {k_{xj}}^{\alpha''} \phi(k_j, .)] {z_{ij}}^{\beta'} {z_j}^{\beta''} \qquad 0 < \alpha', \alpha'', \beta', \beta'' < 1 \tag{38}$$

In (38), the subscript x refers to the commodity (final goods) sector, i to the firm and j to the country. $\phi(.)$ is determined by the contribution to the firm's production from the

28 Handa and Khan (2008) examine Granger causality for the 1960–2002 period by cointegration analysis for a cross-section of thirteen countries at different stages of development, and test four hypotheses: causality is one-way from economic growth to financial development; one-way from financial development to economic growth; two-way between them; or there is no causality of any type. Overall, there is no support for one-way Granger causality from financial development to economic development for any of the countries examined, irrespective of their stage of economic development, but there is support for two-way causality (mainly for developed economies, including the UK, the USA and Japan) or only from economic growth to financial development (mainly for less developed economies). However, their data for the external sources of finance was more limited for the less developed economies.

29 This subsection is illustrative and is not designed to present a complete model.

differential between its knowledge and that associated with the highest capital intensity among the countries of the world.[30] $z_{ij}^{\beta'}$ represents the advantages to the ith firm from using real balances. The firm also benefits from the externality of economy-wide usage of real balances, as represented by the term $z_j^{\beta''}$, so that the social role of money in production technology is explicitly recognized. For the per capita production function of the economy, assuming all firms to be identical, we have:

$$x_j = A_j {k_{xj}}^{\alpha} \phi(.) {z_j}^{\beta} \tag{39}$$

where $\alpha = \alpha' + \alpha''$ and $\beta = \beta' + \beta''$. β can be less than or greater than unity. The productivity of capital is now:

$$\partial x_j / \partial k_{xj} = \alpha A_j {k_{xj}}^{\alpha-1} \phi(.) {z_j}^{\beta} \tag{40}$$

so that the productivity of capital is higher in a monetary than in a barter economy and is also higher if the financial sector has a greater output. Therefore, efficient FI encourages investment in physical capital in the commodity sector. Conversely, the marginal productivity of the services of the financial sector for firms in production is:

$$\partial x_j / \partial z_j = \beta A_j {k_{xj}}^{\alpha} \phi(.) {z_j}^{\beta-1} \tag{41}$$

which is higher in the more capital-intensive economy, thereby providing an incentive for the more capital-intensive economies to use more FI than the less capital-intensive ones. It is also obvious from (38) to (41) that, within a given economy, a sector that has higher values of A and α – so that the productivity of labor and capital are higher in it – would have a higher productivity of real balances and a greater demand for them.

Equations (40) and (41) point to the importance of an efficient financial sector in the economy. The more relevant FI is to the needs of the production sector and the cheaper its provision, the greater will be the amount of it that will be used and the greater will be the return to capital, output per capita, as well as the pre-SS growth rate of output. Whether the SS growth rates will be higher or not is likely to depend upon whether β is less than one or not.[31]

24.12 Investment, financial intermediation and economic development

Most economies seem to show a special connection between increases in capital and efficient FI, leading to higher investment in the commodity sector. We have seen this above in (40), which shows that the marginal productivity of capital increases with FI. But there also seems to be another connection. For this, we have to get away from the concept of *passive investment* – that is, investment responding passively to the saving available for the increase in physical capital – in neoclassical growth models.[32] In the real world, most firms do not possess enough capital through their own resources to finance all the equipment or all the investment that

30 Capital in this context of endogenous growth theories is defined to include both physical and human capital.

31 Note that, to analyse the actual growth rate of output, account will have to taken of the production function and of the labor and capital used in FI.

32 For a Keynesian perspective on money and growth, see Tobin (1965).

they want. Any investment requires additional funds, which can only with difficulty and often at relatively high cost be obtained directly from the ultimate savers in the economy without going through some financial intermediary or other. A restraint on the FI process leads to blockages, reductions and inefficiencies in the flow of funds for investment and consequently to a reduction in investment.

For developing economies, McKinnon (1973), among others, has emphasized the importance of FI to the level and efficacy of investment. He argued that developing economies have fragmented markets with significant constraints on borrowing, which reduce the total amount invested and also channel the available funds into less productive uses, so that unregulated and efficient FI is vital to the growth of such economies.To pursue this argument, let:

$$i_x = i_x(s - s_z, z') \qquad \partial i_x/\partial z' > 0 \tag{42}$$[33]

where:

i_x = investment in the commodity sector
z' = increase in z
s_z = saving devoted to expansion of the FI sector.

In keeping with the emphasis of this section on the financial sector, z' is to be viewed not as the creation of new fiat money but as an expansion of the financial sector with the actual services changing as needed for the growing economy. An expansion of z directly encourages investment i_x so that $\partial i_x/\partial z' > 0$, but the saving devoted to the expansion of z reduces i_x. However, there is an improvement in the efficiency of allocation of saving to investment projects and the sectors of the economy, so that the marginal productivity of investment and that of the capital stock in the production function improves, thereby increasing output per capita in the economy and providing higher growth rates.

Therefore, there are two types of effects of FI on capital. One is by increasing its marginal product, as in (40), for given levels of the aggregate capital stock. The other, as in (42), is the direct effect on the volume of investment of the expansion of the FI sector. Many economists believe that these effects imply higher per capita income and pre-SS growth rates for an economy with a more efficient and competitive financial sector relative to another, otherwise identical, economy. It is also likely that they also do so for the steady states of the two economies.

While these are persuasive arguments that an inefficient and over-regulated financial sector can hold back per capita output and its pre-SS growth rate, such analysis does not provide any guidance on whether the expansion of an efficient and competitive financial sector will increase the SS growth rates of its economy. Monetary economics does not offer conclusive judgment on the latter issue.

Conclusions

The empirical regressions of output growth against the growth rate of money supply over very long periods do not show a significant relationship between these variables, nor does the long-run data show a significant relationship between output growth and the inflation rate.

33 Note that this function is a departure from the neoclassical assumption that investment is solely determined by the available saving.

These findings imply the independence of the long-run growth rate of output from that of the money supply and the inflation rate.

As against the above findings, the stylized facts show that monetary economies have substantially more capital and output per worker than barter ones and that more efficient financial intermediation further increases their values. These facts highlight the way we should think of banking and financial intermediation in the context of growth theory. This is not in terms of the quantity of nominal money in the economy but rather of the real services provided by financial intermediation and of the interaction between this sector and the other sectors of the economy.

The standard neoclassical growth theory encompassing fiat money incorporates money only in terms of its nominal quantity and does not necessarily imply a positive impact of money on output, since the introduction of real balances in it reduces the saving available for the growth of physical capital. There are two reasons for this decline in saving: the increase in real balances increases disposable income and thereby the consumption of commodities. Further, the decreased amount of saving has to be partly allocated to this increase in real balances, thereby reducing investment and capital growth.

The essential requirement for the more realistic effects of money on the economy is the recognition of the basic fact that the production functions in a monetary economy are different from those in non-monetary economies: the former have higher marginal products of capital and labor. Money allows specialization, trade, and expansion in the size of firms to an extent that can never occur under barter and in economies with rudimentary monetary systems. Further, with the shift in transactions technology from barter to one with the use of money, the reduction in transactions time causes an increase in the labor supply. In addition, an efficient financial sector allows the accumulation and more efficient allocation of savings to investment projects, and thereby promotes the growth of capital and increases its overall productivity.

Monetary growth theory applicable to modern economies has to get away from its emphasis on fiat (outside) money and focus on the financial sector as a whole. This sector consists not only of the commercial banks; it also includes investment brokers and stock markets, insurance and mortgage companies, pension funds, etc. Although the financial intermediaries use up some of the resources of the economy in intermediate production, the gains from the use of this intermediate good in final production are much more substantial. The optimal production and usage of financial services in the economy requires the efficiency of the financial sector. Further, the technology of this sector evolves with that of the economy generally, with the proportions of the particular financial assets appropriate to the evolving technology of trade and production of the final goods changing with this technology. Hence, the amount of a given monetary asset does not invariably remain a good indicator of the final output of the financial sector.

Developing economies with evolving monetary systems provide a sort of laboratory for examining the effects of finance on production, trade, investment and capital intensity. A large number of empirical studies support the conclusion that an efficient and evolving financial sector makes a major contribution to economic growth and development.[34]

There is a two-way relationship between the financial sector and the other sectors of the economy. The efficiency of exchange and production in the latter depends on the former, innovation in which can exert an independent impact on the rest of the economy.

34 See Levine (1997) for a review of these.

But, conversely, the latter provides some of the inputs for the former, in the form of the types of capital equipment, knowledge and information technology, the education and skills of employees, etc. The efficiency of the financial sector is also dependent, as is true of the other sectors, on regulations, technological change and other aspects of the economic environment.

Finance is a major sector whose services are needed and used by all the other sectors in the economy. In turn, it uses the inputs and knowledge originating in other sectors. There is, therefore, a symbiotic relationship between an efficient and competitive financial sector and the rest of the economy. This makes it difficult to assign causality on the question of whether the efficiency of the financial sector emanates from the other sectors of the economy, or whether a more efficient financial sector contributes to the efficiency of the latter.

Empirical studies in recent years have begun to separate the relationship between the growth rates of output and the money supply (or inflation), and that between the growth of output and the efficiency and development of the financial sector. On the former set of variables, the usual finding is that there is no evidence of a relationship, or that the relationship is not robust or it is slightly negative. On the latter set of variables, the finding is usually that of a significant positive relationship, with two-way causality between financial and economic development. Therefore, the implication of these empirical findings for monetary policy is that it should support the greater efficiency, diversity and development of the private financial sector.

Summary of critical conclusions

- Monetary growth theory, which focuses on the quantity of money rather than on monetary and other financial institutions, tends to assume that the commodity production function is the same in the monetary economy as in the barter one. Under its usual assumptions, this theory does not provide a clear implication that the use of money provides a higher steady-state standard of living than the non-monetary (barter) economy.
- The use of money is an invention with remarkable externalities for the technology of commodity production and exchange: it increases the size of markets, promotes specialization and trade, and shifts the commodity production frontier drastically. It also increases the supply of labor to firms.
- The monetary sector in the modern economy should be seen as merely one component of the financial sector. The appropriate focus of monetary growth theory should be on the efficiency, diversity and structure of the financial sector rather than on variations in the quantity of money.
- Innovations in the financial sector contribute to the growth rate of output in the pre-SS stage and may also increase the steady-state growth rate.
- The link between financial and economic development is such as to imply a high degree of contemporaneous causality between them. In terms of Granger-causality, most studies tend to show one-way statistical causality from economic development to financial development, or two-way causality.

Review and discussion questions

1. In the context of the traditional neoclassical growth theories with money, is the money growth rate neutral for the steady state and the pre-steady states of the economy? Explain your answer.

2. In the context of the endogenous growth theories with money, is the money growth rate neutral for the steady state and the pre-steady states of the economy? Explain your answer.
3. Is there room for considering changes in the structure and efficiency of the financial sector in the traditional neoclassical growth theories with money? If so, how would you modify the standard model to accommodate such changes? What implications would they have for the steady-state and the pre-steady-state growth rates of the economy?
4. "Rapid increases in the money stock cause inflation and inflation raises interest rates, so that investment is reduced by the rapid increases in the money stock. Consequently, high rates of money growth reduce the growth rate of the economy." Discuss this statement in the context of monetary growth theory. Is there any place for this line of reasoning in the traditional and the endogenous growth theories? Why, or why not? If there is not, does it reflect a deficiency in monetary growth theory?
5. "Given that saving is positively related to the rate of interest and that an increase in the rate of inflation will increase the rate of interest, both saving and the growth rate of the economy will be positively related to the inflation rate and the money growth rate." Discuss in the context of monetary growth theories.
6. "It is vital for understanding the contribution of money to growth that we distinguish between the quantity of money and the size and efficiency of the financial sector." Discuss.
7. Should a distinction be drawn between the growth rate of fiat money provided by the central bank and that of inside money provided by private commercial banks for the proper analysis of money in growth theory? Explain your answer.
8. What is the empirical evidence on the contribution of (a) the quantity of money, (b) the financial sector, to growth? Cite at least one study on each of these. Compare their findings and explain the reasons for any findings that may be different.
9. How would you test for the contribution of (a) the quantity of money, (b) the financial sector, to growth for a selected group of countries? Specify your variables, estimating equations and the frequency of data you would use. Perform this test and interpret your results.
10. "Central bank policies that produce low inflation increase long-run economic growth." Present the relevant theories and empirical evidence on this proposition.
11. LDCs have at times resorted to high money growth rates to finance their development efforts in an attempt to push up their growth rates. What is the theoretical analysis to support this policy? Was it misguided in retrospect and, if so, why?
12. "Banking inefficiency and the excessive regulation of the banking sector and interest rates in some of the LDCs have proved to be highly detrimental to their growth rates." Can this be explained by monetary growth theories? Cite the relevant literature on this subject and relate it to the arguments of the monetary growth theories.

References

Barro, R. "Inflation and economic growth." *Federal Reserve Bank of St. Louis Review*, 78, 1996, pp. 153–69.

Beck, T., and Levine, R. "Stock markets, banks and growth: panel evidence." *NBER Working Paper* no. 9082, 2002.

Bruno, M., and Easterly, W. "Inflation and growth: in search of a stable relationship." *Federal Reserve Bank of St. Louis Review*, 78, 1996, pp. 139–51.

Cameron, R. *Banking in the Early Stages of Industrialization*. New York: Oxford University Press, 1967.

Chari, V.V., Jones, L.E. and Manuelli, R.E. "The growth effects of monetary policy." *Federal Reserve Bank of Minneapolis Quarterly Review*, 19, 1995, pp. 18–32.

Demetriades, P., and Hussein, K. "Does financial development cause economic growth? Time-series evidence from 16 countries." *Journal of Development Economics*, 51, 1996, pp. 387–411.

Demirguc-Kunt, A., and Levine, R. *Financial Structure and Economic Growth*. Cambridge, MA: MIT Press, 2001, pp. 81–141.

Diamond, D.W. "Liquidity, banks and financial markets." *Journal of Political Economy*, 105, 1997, pp. 928–56.

Greenwood, J., and Smith, B. "Financial markets in development and the development of financial markets." *Journal of Economic Dynamics and Control*, 21, 1997, pp. 145–81.

Handa, J., and Khan, S.R. "Financial development and economic growth: a symbiotic relationship." *Applied Financial Economics*, 18, 2008, pp. 1033–49.

Johnson, H.G. "Money in a neoclassical one sector growth model." In H. Johnson, *Essays in Monetary Economics*. London: Allen and Unwin, 1967.

King, R., and Levine, R. "Finance and growth, Schumpeter might be right." *Quarterly Journal of Economics*, 108 (3), 1993, pp. 717–37.

Levhari, D., and Patinkin, D. "The role of money in a simple growth model." *American Economic Review*, 55, 1968, pp. 713–53.

Levine, R. "Financial development and economic growth: view and agenda." *Journal of Economic Literature*, 35, 1997, pp. 688–726.

Levine, R. "Stock market liquidity and economic growth: theory and evidence." *Finance, Research, Education and Growth*, 2003, pp. 1–24.

Levine, R., and Renelt, D. "A sensitivity analysis of cross-country growth regressions." *American Economic Review*, 82, 1992, pp. 942–63.

Levine, R., and Zervos, S. "Stock markets, banks and economic growth." *American Economic Review Papers and Proceedings*, 88, 1998, pp. 537–58.

Lucas, R.E. "On the mechanics of economic development." *Journal of Monetary Economics*, 22, 1988, pp. 3–42.

Lucas, R.E., Jr. "Nobel lecture: monetary neutrality." *Journal of Political Economy*, 104, 1996, pp. 661–82.

McCandless, G.T., Jr, and Weber, W.E. "Some monetary facts." *Federal Reserve Bank of Minneapolis Quarterly Review*, 19, 1995, pp. 2–11.

McKinnon, R.I. *Money and Capital in Economic Development*. Washington, DC: The Brookings Institution, 1973.

Macri, J., and Sinha, D. "Financial development and economic growth: the case for eight Asian countries." *Economia Internazionale*, 55, 2001, pp. 219–37.

Metzler, A.H. "Money, intermediation and growth." *Journal of Economic Literature*, 7, 1969, pp. 27–56.

Page, S., ed. *Monetary Policy in Developing Countries*. London: Routledge, 1993, Ch. 16.

Radford, R.A. "The economic organisation of a P.O.W. camp." *Economica*, 12, 1945, pp. 189–201.

Rajan, R.G., and Zingales, L. "Financial dependence and growth." *American Economic Review*, 88, 1998, pp. 559–86.

Robinson, J. *The Rate of Interest and Other Essays*. London: MacMillan, 1952.

Rousseau, P.L. "Historical perspectives on financial development and economic growth." *Federal Reserve Bank of St. Louis Review*, 85, 2003, pp. 81–105.

Schumpeter, J.A. *The Theory of Economic Development*. Cambridge, MA: Harvard University Press, 1934.

Sidrauski, M. "Rational growth and patterns of growth in a monetary economy." *American Economic Review*, 57, 1967, pp. 534–44.
Solow, R.M. "A contribution to the theory of economic growth." *Quarterly Journal of Economics*, 70, February 1956, pp. 65–94.
Temple, J. "The new growth evidence." *Journal of Economic Literature*, 37, 1999, pp. 112–56.
Tobin, J. "Money and economic growth." *Econometrica*, 33, 1965, pp. 671–84.

Index